DOCUMENT RESUME

ED 042 148 AL 002 507

AUTHOR	Wexler, Kenneth Norman
TITLE	An Automaton Analysis of the Learning of a Miniature System of Japanese. Psychology Series.
INSTITUTION	Stanford Univ., Calif. Inst. for Mathematical Studies in Social Science.
SPONS AGENCY	Edward E. Ford Foundation, New York, N.Y.; National Science Foundation, Washington, D.C.
REPORT NO	TR-156
PUB DATE	24 Jul 70
NOTE	121p.
EDRS PRICE	EDRS Price MF-$0.50 HC-$6.15
DESCRIPTORS	Experiments, Japanese, Mathematical Applications, *Models, *Psycholinguistics, *Second Language Learning, *Semantics, *Syntax
IDENTIFIERS	Automata Theory, *Automaton Analysis

ABSTRACT

The purpose of the study reported here was to do an automata-theoretical and experimental investigation of the learning of the syntax and semantics of a second natural language. The main thrust of the work was to ask what kind of automaton a person can become. Various kinds of automata were considered, predictions were made from them, and these predictions were then tested against data from a learning experiment in order to distinguish between the models. Experimental material was a sub-domain of the set of arithmetic sentence in Japanese, because it was felt that work with a small limited system of language would enable the formulation of precise theories capable of being tested precisely. Syntax learning was felt to be the most important focus of the study; other factors looked for were the influence of semantic practice on syntax learning, and semantic learning. Results of the experiment suggest that a finite automation is not the appropriate representation for the subjects in the experiments; results on semantics suggest that studies of syntax learning that do not include a semantic model may be losing an important component of syntax learning. (Author/FWB)

N-X

AL

ED042148

AN AUTOMATON ANALYSIS OF THE LEARNING OF A MINIATURE SYSTEM OF JAPANESE

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

BY

KENNETH NORMAN WEXLER

TECHNICAL REPORT NO. 156

JULY 24, 1970

PSYCHOLOGY SERIES

AL 002 507

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

TECHNICAL REPORTS

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

(Place of publication shown in parentheses; if published title is different from title of Technical Report, this is also shown in parentheses.)

(For reports no. 1 - 44, see Technical Report no. 125.)

50 R. C. Atkinson and R. C. Calfee. Mathematical learning theory. January 2, 1963. (In B. B. Wolman (Ed.), Scientific Psychology. New York: Basic Books, Inc., 1965. Pp. 254-275)

51 P. Suppes, E. Crothers, and R. Weir. Application of mathematical learning theory and linguistic analysis to vowel phoneme matching in Russian words. December 28, 1962.

52 R. C. Atkinson, R. Calfee, G. Sommer, W. Jeffrey and R. Shoemaker. A test of three models for stimulus compounding with children. January 29, 1963. (J. exp. Psychol., 1964, 67, 52-58)

53 E. Crothers. General Markov models for learning with inter-trial forgetting. April 8, 1963.

54 J. L. Myers and R. C. Atkinson. Choice behavior and reward structure. May 24, 1963. (Journal math. Psychol., 1964, 1, 170-203)

55 R. E. Robinson. A set-theoretical approach to empirical meaningfulness of measurement statements. June 10, 1963.

56 E. Crothers, R. Weir and P. Palmer. The role of transcription in the learning of the orthographic representations of Russian sounds. June 17,. 1963.

57 P. Suppes. Problems of optimization in learning a list of simple items. July 22, 1963. (In Maynard W. Shelly, II and Glenn L. Bryan (Eds.), Human Judgments and Optimality. New York: Wiley. 1964. Pp. 116-126)

58 R. C. Atkinson and E. J. Crothers. Theoretical note: all-or-none learning and intertrial forgetting. July 24, 1963.

59 R. C. Calfee. Long-term behavior of rats under probabilistic reinforcement schedules. October 1, 1963.

60 R. C. Atkinson and E. J. Crothers. Tests of acquisition and retention, axioms for paired-associate learning. October 25, 1963. (A comparison of paired-associate learning models having different acquisition and retention axioms, J. math. Psychol., 1964, 1, 285-315)

61 W. J. McGill and J. Gibbon. The general-gamma distribution and reaction times. November 20, 1963. (J. math. Psychol., 1965, 2, 1-18)

62 M. F. Norman. Incremental learning on random trials. December 9, 1963. (J. math. Psychol., 1964, 1, 336-351)

63 P. Suppes. The development of mathematical concepts in children. February 25, 1964. (On the behavioral foundations of mathematical concepts. Monographs of the Society for Research in Child Development, 1965, 30, 60-96)

64 P. Suppes. Mathematical concept formation in children. April 10, 1964. (Amer. Psychologist, 1966, 21, 139-150)

65 R. C. Calfee, R. C. Atkinson, and T. Shelton, Jr. Mathematical models for verbal learning. August 21, 1964. (In N. Wiener and J. P. Schoda (Eds.), Cybernetics of the Nervous System: Progress in Brain Research. Amsterdam, The Netherlands: Elsevier Publishing Co., 1965. Pp. 333-349)

66 L. Keller, M. Cole, C. J. Burke, and W. K. Estes. Paired associate learning with differential rewards. August 20, 1964. (Reward and information values of trial outcomes in paired associate learning. (Psychol. Monogr., 1965, 79, 1-21)

67 M. F. Norman. A probabilistic model for free-responding. December 14, 1964.

68 W. K. Estes and H. A. Taylor. Visual detection in relation to display size and redundancy of critical elements. January 25, 1965, Revised 7-1-65. (Perception and Psychophysics, 1966, 1, 9-16)

69 P. Suppes and J. Donio. Foundations of stimulus-sampling theory for continuous-time processes. February 9, 1965. (J. math. Psychol., 1967, 4, 202-225)

70 R. C. Atkinson and R. A. Kinchla. A learning model for forced-choice detection experiments. February 10, 1965. (Br. J. math stat. Psychol., 1965, 18, 184-206)

71 E. J. Crothers. Presentation orders for items from different categories. March 10, 1965.

72 P. Suppes, G. Groen, and M. Schlag-Rey. Some models for response latency in paired-associates learning. May 5, 1965. (J. math. Psychol., 1966, 3, 99-128)

73 M. V. Levine. The generalization function in the probability learning experiment. June 3, 1965.

74 D. Hansen and T. S. Rodgers. An exploration of psycholinguistic units in initial reading. July 6, 1965.

75 B. C. Arnold. A correlated urn-scheme for a continuum of responses. July 20, 1965.

76 C. Izawa and W. K. Estes. Reinforcement-test sequences in paired-associate learning. August 1, 1965. (Psychol. Reports, 1966, 18, 879-919)

77 S. L. Blehart. Pattern discrimination learning with Rhesus monkeys. September 1, 1965. (Psychol. Reports, 1966, 19, 311-324)

78 J. L. Phillips and R. C. Atkinson. The effects of display size on short-term memory. August 31, 1965.

79 R. C. Atkinson and R. M. Shiffrin. Mathematical models for memory and learning. September 20, 1965.

80 P. Suppes. The psychological foundations of mathematics. October 25, 1965. (Colloques Internationaux du Centre National de la Recherche Scientifique. Editions du Centre National de la Recherche Scientifique. Paris: 1967. Pp. 213-242)

81 P. Suppes. Computer-assisted instruction in the schools: potentialities, problems, prospects. October 29, 1965.

82 R. A. Kinchla, J. Townsend, J. Yellott, Jr., and R. C. Atkinson. Influence of correlated visual cues on auditory signal detection. November 2, 1965. (Perception and Psychophysics, 1966, 1, 67-73)

83 P. Suppes, M. Jerman, and G. Groen. Arithmetic drills and review on a computer-based teletype. November 5, 1965. (Arithmetic Teacher, April 1966, 303-309.

84 P. Suppes and L. Hyman. Concept learning with non-verbal geometrical stimuli. November 15, 1968.

85 P. Holland. A variation on the minimum chi-square test. (J. math. Psychol., 1967, 3, 377-413).

86 P. Suppes. Accelerated program in elementary-school mathematics -- the second year. November 22, 1965. (Psychology in the Schools, 1966, 3, 294-307)

87 P. Lorenzen and F. Binford. Logic as a dialogical game. November 29, 1965.

88 L. Keller, W. J. Thomson, J. R. Tweedy, and R. C. Atkinson. The effects of reinforcement interval on the acquisition of paired-associate responses. December 10, 1965. (J. exp. Psychol., 1967, 73, 268-277)

89 J. I. Yellott, Jr. Some effects on noncontingent success in human probability learning. December 15, 1965.

90 P. Suppes and G. Green. Some counting models for first-grade performance data on simple addition facts. January 14, 1966. (In J. M. Scandura (Ed.), Research in Mathematics Education. Washington, D. C.: NCTM, 1967. Pp. 35-43.

91 P. Suppes. Information processing and choice behavior. January 31, 1966.

92 G. Groen and R. C. Atkinson. Models for optimizing the learning process. February 11, 1966. (Psychol. Bulletin, 1966, 66, 309-320)

93 R. C. Atkinson and D. Hansen. Computer-assisted instruction in initial reading: Stanford project. March 17, 1966. (Reading Research Quarterly, 1966, 2, 5-25)

94 P. Suppes. Probabilistic inference and the concept of total evidence. March 23, 1966. (In J. Hintikka and P. Suppes (Eds.), Aspects of Inductive Logic. Amsterdam: North-Holland Publishing Co., 1966. Pp. 49-65.

95 P. Suppes. The axiomatic method in high-school mathematics. April 12, 1966. (The Role of Axiomatics and Problem Solving in Mathematics. The Conference Board of the Mathematical Sciences, Washington, D. C. Ginn and Co., 1966. Pp. 69-76.

(Continued on inside back cover)

ED042148

AN AUTOMATON ANALYSIS OF THE LEARNING OF A MINIATURE SYSTEM OF JAPANESE

by

Kenneth Norman Wexler

TECHNICAL REPORT NO. 156

July 24, 1970

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Kenneth Norman Wexler TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

PSYCHOLOGY SERIES

Reproduction in Whole or in Part is Permitted for
any Purpose of the United States Government

© 1970 by Kenneth Norman Wexler
All rights reserved
Printed in the United States of America

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ACKNOWLEDGMENTS

The author gratefully acknowledges the frequent advice and encouragement of Professor Patrick Suppes, who served as chairman of the committee for this dissertation. Conversation with Professor Suppes stimulated many of the ideas in this paper. Thanks are due also to Professors Richard C. Atkinson and William K. Estes, who were the other members of the dissertation committee.

Dr. Kazumi Nishioka deserves thanks for the time he graciously volunteered to teach the author about Japanese arithmetic and for appearing on television as the Japanese speaker. Many thanks are due my wife Sherry for helping with the data analysis and for providing encouragement throughout the course of this work.

Work on the dissertation was begun while the author was an NIMH Predoctoral Trainee in Mathematical Psychology.

The research was supported by a partial grant from the Edward E. Ford Foundation and by National Science Foundation Grant NSFGJ-443x.

AN AUTOMATON ANALYSIS OF THE LEARNING OF A MINIATURE SYSTEM OF JAPANESE

Kenneth Norman Wexler
Stanford University
Stanford, California 94305

I. Introduction

The purpose of this study is to do an automata-theoretic and experimental investigation of the learning of the syntax and semantics of a second natural language. Most studies in the psychological literature (e.g., Braine, 1963; Epstein, 1962) that have tried to deal experimentally with the learning of a small segment of language have analyzed only artificial languages. Crothers and Suppes (1967, Chap. 6) analyzed the learning of some Russian syntax by American college students, making predictions based on alternative conceptions of generative grammar, but did not obtain significant differences based on these conceptions.

The main thrust of this work is to ask what kind of an automaton can a person become? Suppes (1968) showed that there is a sense in which the behavior of any finite automaton can be approached in the limit by a stimulus sampling model. However, the thrust of our work was not to construct a model to capture the trial-to-trial changes in learning, but rather to see what kind of automaton a subject could be at a given point of time, that is, what kind of automaton the learner could use to structure information. We considered various kinds of automata, made predictions from them (and perhaps some auxilliary learning assumptions), and then tested these predictions against data from a learning experiment to distinquish between the models.

Another question we wanted to consider is the role of semantics in language learning. There are two questions here. First, what effect does the introduction of semantics have on syntax larning? Miller and Norman (1964) suggested that perhaps semantics has no direct role in syntax learning, that is, it gives no information to the subject which he uses to learn the syntax. Rather, semantics may have only a motivating role. Minsky (1968, p. 20), on the other hand, conceived of semantics playing a very important part in the understanding of syntax; namely, he claimed that semantics restricts the range of syntactic structures that a sentence can have. This latter view suggests that semantics may have the same effect on syntax learning. That is, the introduction of semantics may aid syntax learning by restricting the possible syntactic structures.

A secondary question we wanted to consider that is relevant to semantics is how the semantics itself is learned. We wanted to look at a simple semantic system to see if we could say anything precise about semantics learning. This was necessary, because almost no work has been done on semantics learning. A recent book (Minsky, 1968) contains a number of articles which describe various attempts to introduce semantics into computers. But very little is said about how a computer might learn these systems.

The above discussions put a number of requirements on our choice of experimental materials. The material had to

(a) be drawn from natural language,

(b) have a simple automata structure that we could specify, and

(c) have a simple semantics that we could specify.

These requirements were met by the material we chose, which is a sub-domain of the set of arithmetic sentences in Japanese. Spoken arithmetic in Japanese has a simple syntax that we could specify. The semantics of the system is simply the semantics of arithmetic.

To give some idea of what we mean by the syntax and semantics of the small system of Japanese we studied, let us give an example in English. Consider the two sentences of English spoken arithmetic:

1. What is two plus three?
2. What plus two is three?

First note that the syntax of the two sentences is different. On a simple level, although the words in both sentences are the same, the order of the words is different. But this is not the only difference between the two sentences; the meanings of the sentences also differ. We took as the meaning (or semantics) of such a sentence its correct answer in arithmetic. Thus, denoting meaning by A, we have A (Sentence 1) = 5 and A (Sentence 2) = 1. Clearly, the meaning of these sentences does not depend only on what words they contain, for sentence 1 and sentence 2 contain the same words yet have different meanings. This of course is exactly the same state of affairs as in natural language in general, e.g., "John loves Mary" is (alas) different in meaning from "Mary loves John."

To what extent are we justified in taking the semantics of a question to be its correct answer? The most serious study of semantics has been in logic, where models which allow one to determine the truth of a sentence are studied. The sentences considered are generally propositions, not questions. However, we can consider a question to be derived by a transformation from a proposition with a variable in it, and we can then say that the meaning of a question is that word or phrase (in our case number) which makes the underlying proposition true with respect to the semantic model.

These considerations are discussed more precisely in Section III. Since they are not central to the major reason for our formulation of the experiment, we will not discuss them further. Before we turn to a brief description of the experiment we want to point out the obvious fact that the experiment deals with only a very small, limited system of language. While our ultimate goal, of course, is to understand the course of language acquisition in general, we have chosen to work with a small experiment so that we can formulate precise theories which are also precisely testable. The rich nature of spoken language in even a young child makes precise testing of, for example, automata models very difficult if they are to apply to the whole range of language. For example, one of the main points of our study is the comparison of two automata, both of which predict learning at asymptote. Discrimination between the automata is possible by comparing details of learning. If we were to apply the same procedure to a large range of natural language, we would first have to write automata to describe this language. We prefer to leave this task as an exercise for the linguists. Then we would have to precisely observe the course of language acquisition. Although a number of investigators have studied, say, child language at a few given points of time, very little of a systematic nature has been said about the course of development over time.

For reasons of the above sort we settled on a simple experiment as an appropriate way to study some aspects of language learning. The materials learned have properties which are sufficiently similar to those demanded by linguistic theory that we call them a "miniature linguistic system." In the experiment subjects learned syntax by being exposed to sentences of this system. We did not teach them any rules. There seems to be general agreement that rules are not directly taught to children learning a first language. Also, it is commonly said that the best way to learn a second language is to go to a country where that language is spoken and learn it, not by learning rules, but by being exposed to the language. For these reasons we did not present rules to the subjects. In general, we feel that the experimental situation provides a reasonable model of some (though certainly not all) of the conditions of language learning. This is especially true of the sentences that are presented with associated "meanings."

The Experiment

A complete description of the experimental method appears in Section IV. Here we give only a brief over-view. Before specifying exactly the set of Japanese sentences used in our experiment, I might mention briefly a pilot experiment in which we used a much broader range of sentences and a different experimental method. The materials were sentences that contained two numbers and a variable and the four operations: addition, subtraction, multiplication and division. The base sentences, in other words, were of the form $x + 2 = 3$, $2 + x = 3$, $2 + 3 = x$, plus the same sentences with the other three operations instead of addition. The integers 0 to 9 were used and only sentences whose correct answer was positive or zero. The sentences were the Japanese sentences derived transformationally from the above equations (see Section III). A subject, an American college student, heard a large number of these sentences, that is, he saw a Japanese speaker saying them on television, and after each sentence saw the correct answer appear on the screen. The subject's job was to write the correct answer in the few seconds provided between the time he heard the sentence and the answer was presented. To give an example, using English instead of Japanese, the speaker might say, "what plus two equals five?" A few seconds transpired, and the digit 3 was flashed on the screen. The subject, who was told the correct answer was a digit from 0 to 9, tried to write the correct answer in the time before it appeared on the screen. The next stimulus might be "what is 6 divided by 3?," and the presented answer would be "2." The only Japanese the subjects knew before these sentences were started were the integers from 0 to 9, which they learned as paired associates. The stimuli were spoken Japanese words and the responses were written numerals.

Subjects did not learn in this experiment. After eight experimental sessions of about 45 minutes each, no subject had yet learned, and it did not look as if they would. The proportion correct did go up over days, but analysis of the results suggested that this was mostly because subjects were guessing better, that is, their answers were drawn from the possible set of answers given the four operations and the two integers they heard. For example, with the two integers 2 and 3 in a sentence, the only answers could be 5, 1, or 6, since 3 divided by 2 is not an integer, and the subjects knew that the answers were integers. So here the subjects learned to guess 5, 1, or 6 on a sentence which contained the integers 2 and 3.

Since the subjects did not learn the structure on this experiment, there is little interesting to say about it, and I shall not discuss it in any more detail. The experiment was useful, however, for we saw how to modify it to obtain more interesting results. First, it seemed that the material was too complex. The sentences we used made up a large portion of the (short) sentences obtained by our grammar. Since this was too much, it was decided that the material in the main experiment was to be limited to the use of one operation, addition. Second, for the same reasons, we used only the integers from 0 to 5. Third, the method of presentation was so difficult that the subjects had no chance to attend to the structure. A new method was adopted which allowed the subject to concentrate on one word at a time. Fourth, since the subjects did not learn, and a number of them complained that they could not tell what the words were (i.e., they could not segment), in the new experiment pretraining was given on the "function" words, i.e., the non-numerical words. This was not translation training, but it was enough to allow the subjects to identify the words when they heard them in sentences.

The Japanese sentences finally selected contained only the addition operation. Each sentence contained exactly two number words and a variable. That is, they were the kind of sentence whose meaning was the answer. They were the Japanese sentences whose base sentences (see Section III) were of the form $x + N = N$, $N + x = N$, or $N + N = x$, where N was an integer from 0 to 5. In Japanese these sentences read, respectively, "ikutsu tasu N wa N desuka," "N tasu ikutsu wa N desuka," and "N tasu N wa ikutsu desuka," where we allowed N to stand for any integer. "Ikutsu" means "what." "Tasu" is Japanese for "add." "Wa" is a post-position, analagous syntactically to English prepositions, but occurs after a noun. "Desu" means "is" and "ka" is a question marker.

Thus, a typical sentence our subjects might hear was, "Ichi tasu ikutsu wa san desuka" ("ichi" is 1 and "san" is 3), "one plus what is three?"

The experiment was carried out in four parts, one part taking place after the previous part was completed. In Part I, the subjects had pretraining on the four function words, "ikutsu," "tasu," "wa," and "desuka." They had to write the first letter of the word when they heard the word spoken by a Japanese speaker on closed-circuit television. Part II consisted of paired-associate training on the Japanese integers from 0 to 5. The speaker said an integer, the subject wrote a digit, and then the correct answer appeared.

Part III was the main part of the experiment, for which Parts I and II were necessary pretraining. Here the subject had to learn the syntax of the addition sentences described above. A sentence was presented slowly. That is, there were a few seconds between each word in the sentence. In this time the subject was to write what words he thought could possibly appear as the next word. This procedure was chosen so as to help the subject learn the syntax and forced him to pay attention to the sentence structure. In the sentences chosen it was always the case that either one or two words could have been the next word. (Subjects were told not to distinguish between numbers, but if they thought a number could be next to simply write N for number). The first position in all sentences could be "ikutsu" or a numeral. The second position was always "tasu." The third position could be "ikutsu" or a numeral, if the first position was a numeral, but if the first position was "ikutsu" then the third position had to be a numeral. In the third position we see for the first time the influence of the history of the sentence (i.e., preceding words). The fourth word had to be "wa." The fifth word had to be a numeral if an "ikutsu" had already appeared. Otherwise it had to be "ikutsu." Once again the influence of the past history of the sentence is seen. The sixth and final word had to be "desuka."

After the sentence was spoken slowly in this manner, it was spoken again, this time at a more normal rate. At this point we considered two groups of subjects. The semantics group (group S) now had the task of answering the Japanese question they had just heard. The sentence was repeated so that the subjects would not have to remember the semantics while concentrating on the syntax in the first part. After the sentence was read for the second time, the subjects wrote a numeral from 0 to 9 which was supposed to be the answer to the question. After the answer, a digit from 0 to 9, appeared visually on the television screen, the next sentence was presented, slowly.

The other group of subjects did not have this semantic task. Instead they had some other task, or none at all, depending on the subgroup in which they were located. (All of these experimental details are presented in Section IV. If they are not important for the present discussion they will be ignored until then.) For this non-semantic ($\overline{S}$) group, no number appeared on the television screen.

The reason for running group $\overline{S}$ was to observe the effects of semantic practice on the learning of syntax. (Only group S was needed to study semantics learning.) The two hypotheses considered about the effect of semantics on syntax learning appear to make different predictions here. If semantics acts as a motivator only, then we have no reason to expect a differential effect on the six responses. That is, group S should do better than group $\bar{S}$ on all responses in the syntax learning task. If, on the other hand, semantics helps syntax learning by restricting the possible structures, then only those responses on which the semantics actually restricts the possibilities should be helped. These considerations will be discussed somewhat more completely in Section III, which deals with the semantic model.

Part IV of the experiment was carried out as a check on Part III. In this part 50 sentences were presented, half of them "grammatical" (G) and half "ungrammatical" (U). G sentences were sentences of the form presented in Part III. U sentences, with the exception of four sentences which we do not discuss now, contained "ikutsu" twice and only one numeral. The second "ikutsu" occurred where one of the numerals would occur in a G sentence. Otherwise, the U sentences were just like the other sentences. An example of a U sentence is "ikutsu tasu 1 wa ikutsu desuka." The sentences were presented one at a time, and the subjects had a few seconds to write a 1 for grammatical or a 0 for ungrammatical. After the correct answer, a 1 or a 0, appeared on the screen, the next sentence was presented.

Part IV was a check of Part III in the following sense. One of the main things we wanted to find out in Part III was whether the subjects would learn the syntax, in a sense to be defined later. If subjects had learned in Part III, then they should learn Part IV quickly, since the information needed in Part IV was a sub-set of the information needed in Part III. Specifically, subjects who had learned Part III should learn Part IV more quickly than subjects who had not learned Part III.

If the difference in learning was not large, we might believe that subjects who had not learned Part III by our definition had nevertheless learned much of the structure.

A summary of the three things we are looking at in this study is

1. Most importantly, syntax learning,
2. The influence of semantic practice on syntax learning, and
3. Briefly, semantics learning.

II. Automata Theories for Syntax Responses

In this section I shall define various kinds of automata to see how to make predictions from them about syntax learning in the experiment. Definitions of automata will be needed in the course of the theoretical development. These definitions are given where needed, but almost no discussion of them is given, since there are many adequate sources for such discussion.

First, we need the definition of finite automata. In an attempt to keep notation standardized in psychological applications of automata, I shall follow the notation of Suppes (1968), which is in essence that of Rabin and Scott (1959). However, the model of an automaton given there is not quite what we need to model this experiment. The model is appropriate in that it is a recognition device that decides what strings are acceptable, but the only way it does this is to determine whether the string brings the machine to an appropriate final state. The final state does seem to have psychological justification, relating to the end of a sentence. However, people understand sentences as they are spoken and, in general, do not have to wait for the end of a sentence to know that it is nonsense or extremely ungrammatical. Specifically, in our experiment, subjects were called on to respond with the next possible inputs after each input. We could define a process whereby they could do this by projecting into the future and seeing what continuations of the string bring the machine to a final state. However, this does not seem at all to be a reasonable model, especially when another one is available.

The problem is that in the Rabin and Scott definition, the transition function M is a function from the Cartesian product of the set of states and set of inputs. So this transition must be defined for every state, input pair. There is a state-diagram for a finite automaton that describes our language in Figure 1. From each state only a few inputs are accepted. The other inputs could be defined as taking the states to which they do not apply into a "collection" state, which then cycles back to itself with each input and is not a member of the set of final states, so that no such string will be accepted. But if the subject is to make his response on the basis of what inputs can come next in the automaton, there is nothing to prevent him from picking the inputs that go to this "collection" state, unless he makes extensive calculations about what can lead to a final state. This seems unreasonable in the limited amount of time he is given.

Insert Figure 1 about here

A model does exist that captures the properties we want. This is what Ginsburg (1962) calls an "incomplete 1-automaton." We follow Ginsburg's development, using as much as possible the notation of Rabin and Scott. Since the class of languages generated by incomplete 1-automata is equal to the class of languages generated by the automata of Rabin and Scott (Ginsburg, 1962, Lemma 4.7, p. 131), we call our machine a finite automaton. The form of definitions closely follows Suppes (1968).

DEFINITION: *A structure* $\mathfrak{A} = \langle A, \Sigma, M, s_0, F\rangle$ *is a finite (deterministic) automaton if and only if*

(1) A *is a finite nonempty set (the set of states),*

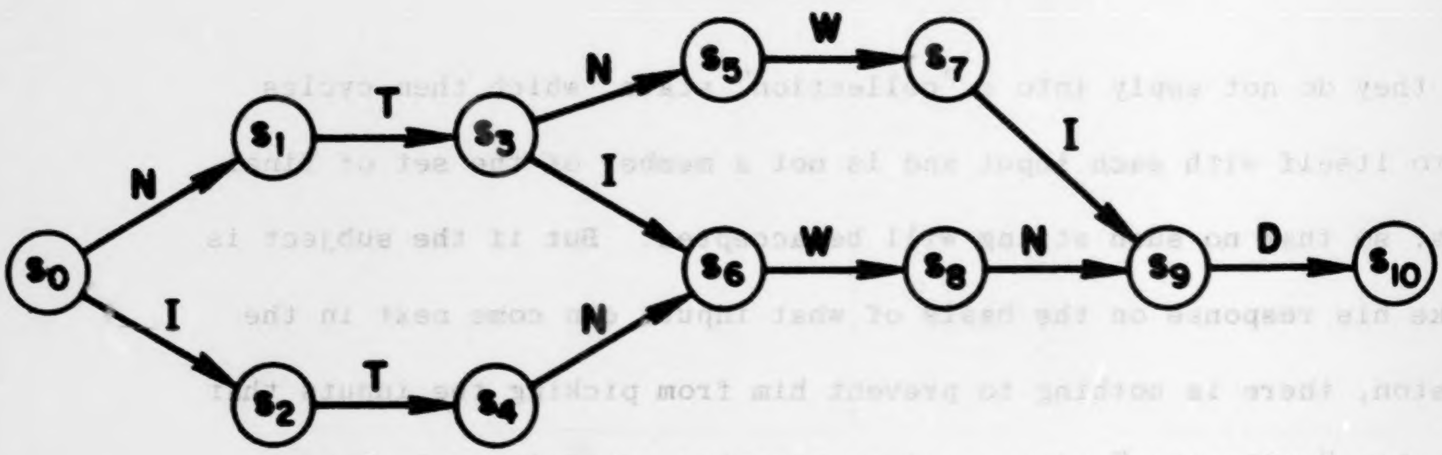
s0
N
I
s1
s2
T
T
s3
s4
N
I
N
s5
s6
W
W
s7
s8
I
N
s9
D
s10

Fig. 1. State-diagram for the finite automaton $\mathcal{J}$.

(2) Σ *is a finite nonempty set (the alphabet or inputs)*,

(3) M *is a function from a subset of the Cartesian product* $A \times \Sigma$ *to* A (*M is the transition table*),

(4) s_0 *is in* A (s_0 *is the initial state*),

(5) F *is a subset of* A (*F is the set of final states*).

The only difference between this definition and that of Suppes is in (3), where the domain of M is specified as a subset of the Cartesian product. Σ^* is the set of finite sequences (strings or tapes) of elements of Σ, including the empty sequence Λ. The function M is extended to a function from a subset of $A \times \Sigma^*$ to A by the following:

DEFINITION: *Let* $\sigma_1, \ldots, \sigma_k$ *be a string in* Σ^* *and let* s *be in* A. $M(s, \sigma_1, \ldots, \sigma_k)$ *is said to exist if each state* $s_1 = s$ *and* $s_{i+1} = M(s_i, \sigma_i)$ *exists, for* $i \leq k$. *When* $M(s, \sigma_1, \ldots, \sigma_k)$ *exists, it is defined to be the state* s_{k+1}.

DEFINITION: *A string* x *of* Σ^* *is accepted by* $\mathfrak{A}$ *if and only if* $M(s_0, x)$ *exists and is in* F. *A string accepted by* $\mathfrak{A}$ *is a sentence of* $\mathfrak{A}$.

DEFINITION: *The language* $T(\mathfrak{A})$ *generated by* $\mathfrak{A}$ *is the set of all strings accepted by* $\mathfrak{A}$.

At this point I want to consider some special definitions that attempt to model what the subject had to do in the syntax learning task. The subject had to decide, according to the instructions, what the next possible words could be, that is, what the next acceptable inputs were. If we conceive of the subject as a finite automaton, we can define a notion of *response* that captures the process of the subject's response.

DEFINITION: *The response* r *of the finite automaton* $\mathfrak{A}$ *is a function from the set of states* A *to the set of subsets of inputs* 2^{Σ} *such that for* $s \in A$,

$$r(s) = \left\{\sigma \in \Sigma \text{ such that there is an } s' \in A \text{ such that } M(s,\sigma) = s'\right\}.$$

In other words, given the state of the automaton, r is the set of possible next inputs. The motivation for defining "response" is that if a subject in our experiment "became" a finite automaton and his task was to write the next possible inputs, he would do so based on his current state, i.e., produce the "response." To give an example, consider the finite automaton $\mathcal{F}$. Figure 1 shows that $M(s_0,N) = s_1$ and $M(s_0,I) = s_2$ and that there is no other input σ such that $M(s_0,\sigma)$ exists. Therefore, by the definition, $r(s_0) = \left\{N,I\right\}$. Likewise, $r(s_1) = \left\{T\right\}$ and $r(s_{10}) = \emptyset$, the empty set. The function r is computed in the same manner for the other states.

Instead of defining the finite automaton as a recognition device and then constructing the "response" of the automaton, we might note another possible approach to modeling our experiment would be to define the automaton as an output device, or a Moore machine. That is, each time the machine reached a state it yielded an output that depended only on the current state. For our purposes the output would play the role of response in the current construction, and no special definition of response would be needed.

A problem with this approach is that an entirely new output function would have to be defined. Let O be the output function and the automaton $\mathcal{F}$, as in Figure 1. Then the output alphabet would be defined as 2^{Σ}, where $\Sigma = \left\{N,I,T,D,W\right\}$ and set, for example, $O(s_0) = \left\{N,I\right\}$ and

$O(s_1) = \{T\}$. In other words, it is clear that we would set $O(s_i)$ $r(s_i)$. By adopting our method we have actually defined an output. The important point is that a natural method has been provided for finding the output instead of arbitrarily assigning the appropriate values.

I now return to our development.

DEFINITION: *A language* L *is a sub-set of* Σ^*. *An initial segment* z *of* L *is a string* $z \in L$ *such that there is a string* $w \in (\Sigma^* - \{\Lambda\})$ *such that* $zw \in L$.

The elements of L have been excluded from being initial segments, because this is useful for experimental purposes. For other purposes, it might be desirable to include them. Denote by $\mathscr{I}(L)$ the set of all initial segments of L.

DEFINITION: *The next-word function of* L *is a function* n *from* $\mathscr{I}(L)$ *to* $(2^\Sigma - \{\Lambda\})$ *such that if* $w \in \mathscr{I}(L)$ (*i.e.*, w *is an initial segment of* L), *then* $n(w) = \{\sigma \in \Sigma$ *such that there is a* $z \in \Sigma^*$ *such that* $w\sigma z \in L\}$. In other words, given an initial string, n tells us what letters may come next. Note that in the above definition $n(\Lambda)$ is the set of initial letters of L.

DEFINITION: *Let* $\mathfrak{A}$ *be an automaton and* L *be a language. We say that* $\mathfrak{A}$ *responds correctly to* L *if (letting* r *be the response of* $\mathfrak{A}$, *and* n *be the next-word function of* L),

(1) $r(s_0) - n(\Lambda)$,

and for all $x \in (\mathscr{I}(L) - \{\Lambda\})$

(2) $M(s_0,x)$ *exists and* $r(M(s_0,x)) = n(x)$.

This definition explains what we mean by learning syntax. A subject who learns the syntax will "respond correctly." That is, he will give the appropriate next possible words.

Now, consider an automaton for the sample of Japanese arithmetic. As mentioned in Section I, we do not have to consider sentences that differ only in the numerals they use. We can assume that, due to our instructions, the subject codes a numeral he hears as N. At any rate, the responses contain no individual numeral, only N, and our theory aims to explain the responses. Using the notation of the first letter of a Japanese word to stand for that word and using N to stand for a numeral, there are exactly three sentences in our language, which we will call J.

$$J = \{NTNWID, NTIWND, ITNWND\}.$$

A transition table for a finite automaton $\mathcal{J}$ such that $T(\mathcal{J}) = J$ is shown in Figure 2. This automaton has the state-diagram shown in Figure 1.

Insert Figure 2 about here

$\mathcal{J} = \langle A,\Sigma,M,s_0,F\rangle$ where $A = \{s_i, 0 \le i \le 10\}$, $\Sigma = \{N,I,T,W,D\}$ and $F = \{s_{10}\}$. A simple calculation shows that $\mathcal{J}$ responds correctly to J, and that $T(\mathcal{J}) = J$. Therefore, if we assume that our learner becomes a finite automaton, we would predict that in the limit he will learn the syntax of J in the sense that he will respond correctly to J.

However, the intuitive feel of the automaton $\mathcal{J}$ is not quite right. The states do not seem to make psychological sense. For example, after one input, $\mathcal{J}$ is in either s_1 or s_2 but s_1 and s_2 can both accept T. Since in both sentences T appears at the same time, somehow the states that accept them should be related. In other words, if an input word appears in the same place in two different sentences, it should show up in the state structure of the automaton. The next

	N	I	T	W	D
0	1	2			
1			3		
2			4		
3	5	6			
4	6				
5				7	
6				8	
7		9			
8	9				
9					10
10					

Fig. 2. Transition table for the finite automaton $\mathcal{J}$. For simplicity, a state is denoted i instead of s_i as in the text.

definition yields a kind of finite automaton that seems to have the properties we want.

DEFINITION: An ordered-state finite automaton (OSA) is a finite automaton such that for all i,j,k if $M(s_i,\sigma_j)$ and $M(s_i,\sigma_k)$ exist, then

$$M(s_i,\sigma_j) = M(s_i,\sigma_k).$$

It follows from the definition that at any given time (i.e., after a given number of inputs) there is only one state that an ordered-state automaton can be in, no matter what the past history. The ordered-state automaton is of interest to us mainly where there are transitions that are not defined. In an ordered-state automaton, if all transitions are defined, that is, if $M(s,\sigma)$ exists for all states s and inputs σ, then clearly for any integer k either all strings of length k are accepted or none are. That is, whether the automaton accepts or rejects a string depends only on its length.

To us it makes a lot of intuitive sense to suppose a subject becomes a sequential automaton. The state of the automaton is directly linked to time. The subject can learn where T appears, in a sense, by learning that it always appears in second position. What sequential automaton can behave like J? None, as shown by the following.

THEOREM: There is no ordered-state automaton that responds correctly to J.

Proof: Suppose $\mathfrak{A}$ is an OSA which responds correctly to J. Recall $J = \{NTNWID, NTIWND, ITNWND\}$. Since $\mathfrak{A}$ responds correctly to J, $M(s_0,I)$ and $M(s_0,N)$ exist. Therefore, by the definition of an OSA, $M(s_0,I) = M(s_0,N)$. Call this state s_1, and set $M(s_1,T) = s_2$. Therefore $M(s_0,NT) = s_2$ and $M(s_0,IT) = s_2$. Therefore letting n be the next-word function of J and r be the response of $\mathfrak{A}$),

$r(s_2) = n(NT)$ and $r(s_2) = n(IT)$. Therefore $n(NT) = n(IT)$.

Inspection of J reveals that

$n(NT) = \{N, I\}$ and $n(IT) = \{N\}$. Therefore $n(NT) \neq n(IT)$.

Therefore we have a contradiction, and the theorem is proved.

Since there is no ordered-state automaton that responds correctly to J, we can predict that if the subject becomes an OSA, he won't learn the syntax of J, that is, he won't respond correctly at asymptote. In fact, this was the first hypothesis we developed about the experiment.

It is interesting that we can predict the subject will not learn from assuming that he becomes an ordered-state automaton independently of any assumptions about the course of learning, that is, of the trial-by-trial changes in the subject's responses or even his automaton. The prediction rests upon the way the subject structures information. An ordered-state automaton severely limits this structure. If we add the additional assumption that the automaton is loop-free, it is clear that the language generated by an OSA must be of the form $A_1A_2 \ldots A_n$ (where $A_i \in 2^{\Sigma}$) in the language of regular expressions, or, in other words, a Cartesian product of sets of inputs. Of course, there is no such representation for J.

If the subject ignores the past sequence of words, except for letting them tell him at what point of time the input is, his responses will be simply those words which can come at the next point of time for some input string, and his sequence of responses will be NI, T, NI, W, NI, D where NI indicates that both N and I were placed in a box. This, of course, can be cast in the Cartesian product form and be generated by an ordered-state automaton. Indeed, this is the automaton one would expect to find in the subject's responses if he becomes an OSA. A state diagram for such an automaton, $\mathcal{A}$, appears in Figure 3.

Insert Figure 3 about here

Although we have shown that no ordered-state finite automaton responds correctly to J, it is still possible that in some sense an OSA might respond correctly to J in the limit, with probability arbitrarily close to 1, so in practice we could not rule out such a machine. To investigate this possibility, we make the following:

DEFINITION: The probabilistic response r of the finite automaton $\mathfrak{A}$ is a set of random variables $r(s)$ for each state s of $\mathfrak{A}$, taking values in 2^{Σ}. The automaton, response-pair $(\mathfrak{A}, r)$, responds correctly up to ϵ to a language L (for $\epsilon > 0$) if

(1) $\Pr(r(s_0) = n(\Lambda)) > 1 - \epsilon$,

and for all $x \in \mathscr{C}(L) - \{\Lambda\}$,

(2) $M(s_0,x)$ exists and $\Pr(r(M(s_0,x)) = n(x)) > 1 - \epsilon$.

DEFINITION: Let $(\mathfrak{A}_i, r_i)$, $i = 1,2,\ldots$, be a sequence of pairs of finite automata and probabilistic responses for the automata. We say the sequence can respond correctly with probability 1 to L if, for all $\epsilon > 0$, there is an integer N (depending on ϵ) such that $(\mathfrak{A}_N, r_N)$ responds correctly up to ϵ to L.

In this last definition we could have made an even stronger condition, namely, we could have required some kind of convergence, that is, in some sense, later automata in the sequence get closer to responding correctly. This would be in line with the usual convergence to

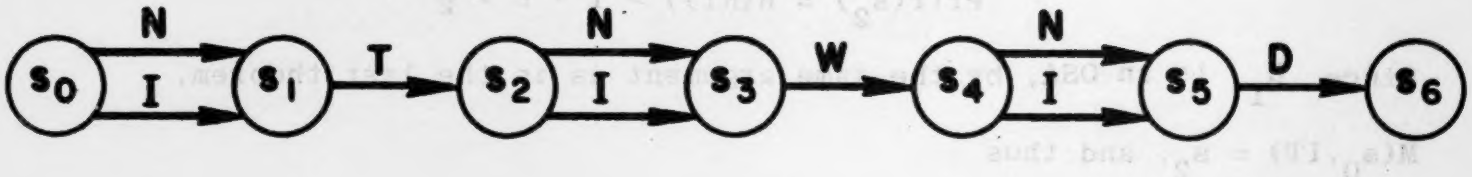
s0
N
I
s1
T
s2
N
I
s3
W
s4
N
I
s5
D
s6

Fig. 3. State-diagram for an ordered-state finite automaton $\mathscr{A}$.

probability 1 definitions. But we have stayed with this weaker condition, because as we shall see now, ordered-state automata cannot even meet the weak condition.

THEOREM: *There is no sequence* $(\mathfrak{A}, r)$ *of automaton, response pairs* $(\mathfrak{A}_i, r_i)$ *where each* $\mathfrak{A}_i$ *is an ordered-state automaton such that* $(\mathfrak{A}, r)$ *can respond correctly with probability* 1 *to* J.

Proof: Suppose $(\mathfrak{A}, r)$ is such a sequence. Pick $\epsilon < \frac{1}{2}$ and let $(\mathfrak{A}_i, r_i)$ respond correctly up to ϵ to J. The proof is similar to that for the deterministic theorem. For this automaton, response pair, since the automaton responds correctly to J, $M(s_0, NT) = s_2$ exists, and

$$Pr(r(s_2) = n(NT)) > 1 - \epsilon > \tfrac{1}{2} .$$

Since $\mathfrak{A}_i$ is an OSA, by the same argument as in the last theorem, $M(s_0, IT) = s_2$, and thus

$$Pr(r(s_2) = n(IT)) > 1 - \epsilon > \tfrac{1}{2}.$$

But inspection of J reveals that

$$n(NT) = \{N, I\} \quad \text{and} \quad n(IT) = \{N\} .$$ Therefore,

$Pr(r(s_2) = \{N, I\}) > \frac{1}{2}$ and $Pr(r(s_2) = \{N\}) > \frac{1}{2}$, which is a contradiction, and the theorem is proved.

This last theorem is rather strong in regard to the capabilities of ordered-state finite automata. No matter how we might try to approach J with an OSA, changing both the automaton and the response distribution, there is no chance of coming close to responding correctly to J.

If the subject does learn, though, that is, if he responds correctly at asymptote, are we forced to conclude that he is a finite automaton of the non-ordered-state type, such as $\mathcal{J}$? Somehow we would like to find an automaton that preserved the ordered-state property while using the past appropriately. These properties can be found in an appropriate push-down store (PDS) automaton (first called so by Newell, Shaw and Simon, 1959). Two mathematical treatments, which differ slightly, may be found in Chomsky (1963) and Ginsburg (1966). However, we do not need anything like the full power of the PDS automata. We are not introducing the PDS to obtain more generative power, since the finite automata are strong enough in this respect; rather we introduce PDS automata in order to obtain different kinds of structure. In particular, we will not need the PDS ability to erase from memory. What we have is the same structure of a special case of what Chomsky called a "transducer," but we do not consider the machine as a mapping from inputs into memory strings as a transducer does. The essential structure is the same, because neither a transducer nor our machine allows erasures, and thus, neither allows past memory to be inspected by the machine. For our purposes we only need one element in memory at any given time, and this again is different from a general PDS. Our machine is also deterministic. As far as we know, an automaton exactly like ours has not been defined in the literature. As far as possible we will try to make our definition a special case of Ginsburg's (1966, p. 59). This, however, is not completely possible because, for the same reasons we gave for the finite automaton definition, we want the transition function to be defined on only a *subset* of the appropriate Cartesian product, whereas Ginsburg defines the function on the full set. Nevertheless, these notions can be defined in a manner similar to that for finite automata.

DEFINITION: *A structure* $\mathfrak{A} = \langle A, \Sigma, \Gamma, M, z_0, s_0, F, e \rangle$ *is a 1-memory store (1-MS) if and only if*

(1) A *is a nonempty finite set (states),*

(2) Σ *is a nonempty finite set (inputs),*

(3) Γ *is a nonempty finite set* (memory elements),

(4) M *is a function from a subset of* $A \times \Sigma \times (\Gamma \cup \{e\})$ *to* $A \times (\Gamma \cup \{e\})$ (M *is the transition table) such that*

a) *if* $M(s,\sigma,m)$ *exists, then* $M(s,\sigma,m) = (s',e)$ *if and only if* $m = e$

b) *if* $M(s,\sigma,e)$ *exists then there is no* $m \in \Gamma$ *such that* $M(s,\sigma,m)$ *exists (the deterministic condition),*

(5) z_0 *is an element of* Γ (z_0 *is the start push-down symbol),*

(6) s_0 *is in* A (s_0 *is the start state),*

(7) F *is a subset of* A (F *is the set of final states),*

(8) e *is not in* Γ (e *is the empty memory element).*

Actually there is little difference between the foregoing definition and the usual PDS definition. What makes our machine a "1-memory store" is the manner in which it moves. The way we conceive of the 1-MS as moving is the following. The machine is in a state, has one memory element at that time, receives an input, and as a result of those three properties, switches to another state, and changes the memory element to another one. In order to realize this process we define the following function M' from a subset of $A \times \Sigma^* \times \Gamma$ to $A \times \Gamma$.

DEFINITION: *Let* $\sigma_1 \ldots \sigma_k$ *be a string in* Σ^*, *and let* s *in* A *and* m *in* Γ. $M'(s,\sigma_1\ldots\sigma_k,m)$ *is said to exist if there is a sequence of states in* A, $s_1, \ldots, s_k$, *where* $s_1 = s$, *and a sequence of memory elements in* Γ, $m_1, \ldots, m_{k+1}$, *and* $m_1 = m$, *such that for* $i \leq k$, *either*

(1) $(s_{i+1},m_{i+1}) = M(s_i,\sigma_i,m_i)$ *exists, or*

(2) $(s_{i+1},e) = M(s_i,\sigma_i,e)$ *exists and* $m_i = m_{i+1}$.

When $M'(s,\sigma_1 \ldots \sigma_k,m)$ exists, it is defined as (s_{k+1},m_{k+1}). We have defined the function as M' instead of M, inasmuch as it is not quite an extension of M, because when $k = 1$ we have $M(s,\sigma_1,m) \neq M'(s,\sigma_1,m)$ when $M(s,\sigma_1,e)$ exists. Since it will not cause confusion, from now on we will call this function M instead of M'.

Now we can see what e does in the definition of a 1-memory store. When a transition $M(s,\sigma,e) = (s',e)$ exists, it means that when a 1-MS is in state s, has memory element m, and receives input σ, it switches to state s' and leaves the memory element unchanged. Of course, given our definition of a 1-MS, we could have accomplished the same result by writing out such a rule for each memory element. But there are structural reasons for not doing this. In our discussion of J we will see that the subject operates sometimes as if he is ignoring what is in memory. The deterministic condition insures that the 1-MS is never confused and has at most one rule to follow. This condition is similar to a condition in Ginsburg's (1966, p. 74) definition of a deterministic push-down automaton, but it does not make Ginsburg's assumption that it is always possible to make a next move.

DEFINITION: A string x of Σ^* is accepted by a 1-MS $\mathfrak{A}$ if and only if $M(s_0,x,z_0)$ exists and is in F. The language $T(\mathfrak{A})$ generated by $\mathfrak{A}$ is the set of all strings accepted by $\mathfrak{A}$.

It is easy to show that the class of languages generated by 1-memory stores is equal to the class of languages generated by finite automata. In general we need fewer states for a 1-memory store than for the equivalent finite automaton. For any finite automaton we can find an equivalent 1-MS with the same number of states simply by adding a memory element which has no effect. In general, however, we can find an equivalent 1-MS with fewer states.

We now have to make our special definitions for modeling our experiment just as we did for finite automata. The definitions will be just like those for finite automata except that, of course, the memory element has to play its natural role.

DEFINITION: *The response* r *of the 1-MS* $\mathfrak{A}$ *is a function from* $A \times \Gamma$ *to* 2^{Σ} *such that for* s *in* A *and* m *in* Γ,

$$r(s,m) = \left\{ \begin{array}{l} \sigma \ \text{such that there is an } s' \text{ in } A \text{ and an } m' \text{ in } \Gamma \cup \{e\} \\ \text{such that } M(s,\sigma,m) = (s',m') \text{ or } M(s,\sigma,e) = (s',m') \end{array} \right\} .$$

We see here another reason why our deterministic condition is necessary, namely, to insure that the response of $\mathfrak{A}$ is not ambiguous.

DEFINITION: *Let* $\mathfrak{A}$ *be a 1-MS and* L *be a language. We say that* $\mathfrak{A}$ *responds correctly to* L *if (letting* r *be the response of* $\mathfrak{A}$*, and* n *be the next-word function of* L),

(1) $r(s_0,z_0) = n(\Lambda)$,

and for all $x \in (\mathcal{A}(L) - \{\Lambda\})$,

(2) $M(s_0,x,z_0)$ *exists and* $r(M(s_0,x,z_0)) = n(x)$.

DEFINITION: *A 1-MS is an ordered-state 1-memory store if for all* s_i *in* A, σ_j *and* σ_k *in* Σ *and* m_p *and* m_q *in* $\Gamma \cup \{e\}$, *if* $M(s_i,\sigma_j,m_p)$ *and* $M(s_i,\sigma_k,m_q)$ *exist, then they are equal.*

A state diagram for a 1-MS $\mathcal{P}$ appears in Figure 4. A triple labeling

Insert Figure 4 about here

a directed line between two states has the obvious interpretation. That is, suppose $s_i \xrightarrow{(\sigma,m,m')} s_j$. Then $M(s_i,\sigma,m) = (s_j,m')$. The 1-MS is

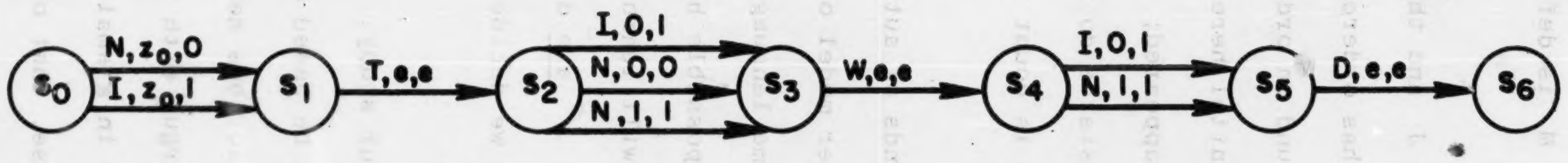
s_0
N, z_0, 0
I, z_0, 1
s_1
T, e, e
s_2
I, 0, 1
N, 0, 0
N, 1, 1
s_3
W, e, e
s_4
I, 0, 1
N, 1, 1
s_5
D, e, e
s_6

Fig. 4. The ordered-state 1-memory store $\mathcal{P}$.

defined as $\mathcal{P} = \langle A, \Sigma, \Gamma, M, z_0, s_0, F, e \rangle$ where $A = \{s_i, 1 \le i \le 6\}$, $\Sigma = \{N, I, T, W, D\}$, $\Gamma = \{0, 1, z_0\}$, $F = \{s_6\}$ and M is defined by the state-diagram in Figure 4.

A simple calculation shows that $T(\mathcal{P}) = J$ and that $\mathcal{P}$ responds correctly to J. It is also clear that $\mathcal{P}$ has ordered states. So, unlike the finite automaton case, we have found an ordered state 1-MS that responds correctly to J. What is essential here is the memory which keeps track of whether an I has yet appeared; it becomes 1 if it has, and 0 if it has not. Therefore the states do not have to keep track of this important history; all they do is count the number of past inputs (i.e., keep track of time).

Now that we have found two different kinds of automata that respond correctly to J, can we tell which is a better model of the subject? Since $\mathcal{F}$ and $\mathcal{P}$ both accept exactly the same language and both respond correctly to J, there is no discrimination possible here. If subjects become either one of the two automata, they will learn, and so we can distinguish them on this basis. Yet $\mathcal{F}$ and $\mathcal{P}$ are different; that is, their structures are different. How can we decide which of the two models is a better one to describe subjects?

This is one of the major questions of our study. The solution to such a question in linguistics usually would be based on introspection, that is, an attempt to decide which model describes mental structure best on the basis of feel. Our point is not to argue with that method, which often is the only one available, but to show in a small example how other kinds of data might be available. In our case, that other kind of data involves learning. If the structures of the two automata are different, very likely this structural difference is reflected in the details of learning, even if both automata predict learning at asymptote.

In order to make predictions about learning, we must make some assumptions about the course of learning, but our assumptions will be reasonably weak (though they may be wrong) and fairly general. It would not be easy to fit a model well to this relatively complex experiment; that is not our goal. In fact, our assumptions will not be strong enough to predict any of the statistics of learning.

A reasonable model for how a subject may become a finite automaton (and learns to respond correctly in the experiment) is the following. After each input the subject is in the correct state. That is, even though he may not have had the appropriate transition function to get to that state, when the input comes in it switches him to that state. This is important because then the subject will have a chance to learn which inputs may be accepted in that state, that is, what the correct response to that state is.

When a subject is in a state and an input comes in, we assume that the subject to some extent learns that that input is part of the correct response to the state. We do this by assuming that there is an increment in the probability that the subject will include the input in his response to that state. We need the following:

DEFINITION: *A pair* (s,σ) *for* s *in* A *and* σ *in* Σ *(for a finite automaton* $\mathfrak{A}$*) appears in a string* x *in* Σ^* *if there are strings* y *and* z *in* Σ^* *such that* $x = y\sigma z$ *and* $M(s_0,y) = s$.

DEFINITION: *Let* $\mathfrak{A}$ *be a finite automaton. For each* s *in* A*, the learner's response to* s *is a random variable* $R(s)$ *taking values in* 2^{Σ}.

DEFINITION: *A presentation schedule is a sequence* $x_1,\ldots,x_i,\ldots$ *of strings in* Σ^*. x_i *is presented on trial* i. *A pair* (s,σ) *is presented on trial* i *if* (s,σ) *appears in* x_i. *The learning sequence for a state* s *is a sequence of learner's responses to* $s, R_0, R_1, \ldots, R_i \ldots$. *Let* f *be a function from* $[0,1]$ to $[0,1]$ *such that* $f(x) > x$ *for* $x < 1$ *and* $f(1) = 1$.

Assumption: For a finite automaton $\mathfrak{A}$, letting $p_i(s,\sigma) = \Pr(\sigma \in R_i(s))$, we assume

$$p_{i+1}(s,\sigma) = \begin{cases} f(p_i(s,\sigma)) & \text{if } (s,\sigma) \text{ was presented on trial } i \\ p_i(s,\sigma), & \text{otherwise.} \end{cases}$$

In other words, if the state and next input were presented on the trial, the subject increases his probability of making the appropriate response. Otherwise he leaves the probability unchanged. If we assume that the initial probability of including an input in a response is 0, then no wrong input will ever be included in a response, and the subject's only problem will be to learn the correct responses. This assumption of the f function is rather general and leaves room for a variety of models, including linear and n-state Markov models. However, the assumption does preclude forgetting, but forgetting could have been included by introducing a forgetting function. The predictions we make would then have turned out in a sense even stronger, and there is no reason to introduce this extra complexity.

The predictions we make from this assmption are applied to the finite automaton $\mathfrak{J}$ (Figure 1). The first prediction involves the inputs T and W. We just consider T here, because the derivation for W is the same. The above considerations lead to the conclusion that, if before trial $i + j + 1$, letting $\#(s_1)$ be the number of presentations of (s_1,T) and $\#(s_2)$ be the number of presentations of (s_2,T), then

$$\Pr(T \in R(s_1) \mid \#(s_1) = i \text{ and } \#(s_2) = j) = \Pr(T \in R(s_1) \mid \#(s_1) = i).$$

In other words, the number of appearances of T does not count when they bring the automaton to state s_2 instead of s_1. The same kind of prediction may be made with W replacing T. The prediction is not made for D, because in $\mathcal{J}$, D only appears with one state. Similar predictions may be tested statistically in a number of ways, and there is no need to discuss them here, since they are discussed in Section V. Essentially the prediction say that there are two kinds of trials on which T appears and that learning T on one does not help on the other.

A variant of this prediction involves comparing learning on, say, the response for the second and third inputs. In the experiment the probabilities of presenting each of the three sentences of J are

$$\Pr(\mathrm{NTNWID}) = \tfrac{1}{2}$$

$$\Pr(\mathrm{NTIWND}) = \tfrac{1}{4}$$

$$\Pr(\mathrm{ITNWND}) = \tfrac{1}{4}.$$

Our assumption leads to the prediction that after i presentations of (s,σ), $p(s,\sigma) = f^{i}(0)$, where the notation f^{i} means function composition of f, i times. So if there are t trials in all,

$$\Pr(N \in R(s_0)) = f^{3/4t}(0) = \Pr(T \in R(s_1)), \text{ and}$$
$$\Pr(I \in R(s_0)) = f^{1/4t}(0) = \Pr(T \in R(s_2)). \qquad (1)$$

One way of interpreting these equations is that the rate of learning for the response predicting the second input should equal that for the response predicting the first input. We test this in our experiment.

By now clearly something general is going on, There is something about the finite automaton model that does not let the same inputs become connected in the appropriate way. Certainly if T appears in the second position in two different sentences it should be learned faster, that is, both kinds of learning events should help each other. The empirical results bear this out.

In order to look at the same predictions for a 1-memory store model we need the

DEFINITION: Let $\mathfrak{A}$ be a 1-MS. For each s in A and m in $\Gamma \cup \{e\}$, the learner's response to (s,m) is a random variable $R(s,m)$ taking values in 2^{Σ}. A triple (s,σ,m), where $m \neq e$, appears in a string x in Σ^* if there are strings y and z in Σ^* such that $x = y\sigma z$ and $M(s_0,y,z_0) = (s,m)$. We say (s,σ,e) appears in x if $x = y\sigma z$ and there is an m in Γ such that $M(s_0,\sigma,z_0) = (s,m)$ and $M(s,\sigma,e)$ exists.

Other definitions are just as before, making the appropriate new definition of "appears."

Assumption: For a 1-MS $\mathfrak{A}$, letting $p_i(s,\sigma,m) = \Pr(\sigma \in R_i(s,m))$, we assume

$$p_{i+1}(s,\sigma,m) = \begin{cases} f(p_i(s,\sigma,m)) & \text{if } (s,\sigma,m) \text{ was presented on } i \\ p_i(s,\sigma,m), & \text{otherwise}. \end{cases}$$

The 1-MS determines the next response using the current state and memory element, and learns in this manner also. What is especially important for us is that it can determine the next response by using e and ignoring memory completely. Thus, when it is ready to accept T, the 1-MS ignores the fact that there is a 1 or a 0 in memory, that is, that the past history is different, and thus, can let each presentation of T help in learning T as one response. This is exactly what the finite automaton model cannot do, as we saw previously. To see the result for the 1-MS model, we consider the cases when the finite automaton $\mathfrak{A}$ is in s_1 or s_2, that is, when an N or I starts the sequence, respectively. Suppose i sequences start with N and j with I, as before. Then by our learning assumption, for the 1-MS (Figure 4), noting that (s_1, T, e) is presented on all of these trials, we have

$$Pr(T \in R(s_1, e)) = f^{i+j}(0).$$

We see that contrary to the result for finite automata, all trials have an effect on the learning of the single T response. This result was found to hold in the experiment and thus helped to suggest that $\mathcal{P}$ is a more appropriate model than $\mathfrak{A}$.

In contrast to the set of equations (1), the 1-MS model predicts

$$\begin{aligned} Pr(N \in R(s_0, z_0)) &= f^{\frac{3}{4}t}(0) \\ Pr(I \in R(s_0, z_0)) &= f^{\frac{1}{4}t}(0) \\ Pr(T \in R(s_1, e)) &= f^{t}(0). \qquad (2) \end{aligned}$$

The first two equations of the set are the same as in (1). But the last one is different from the second two of (1). Equation (2) predicts faster learning for T than for the first response set, in contrast to Equation (1) which predicts equivalent learning. The results bear out the prediction of Equation (2), and the 1-MS model agrees better with data once again.

The essential property of $\mathscr{P}$ that allows us to make these predictions is its ordered-states that can tie together two identical inputs occurring from the same state, but with different histories. The finite automaton cannot do this.

Perhaps a general word is in order. There is a certain sense in saying that part of what we are studying is the psychological process known as "generalization." For example, the 1-memory store model predicts that a subject generalizes from a T with one history to a T with another history and says that in a certain sense they are the same. This generalization takes place over time, but relative time, that is, relative to the place of the word in a sequence, since the two appearances of T are very different in absolute time. The point I am trying to make is that any study of generalization demands a structural model of some kind. Traditional generalization studies have been done in areas where the generalization operated over a simple structure, namely, one continuous dimension such as the frequency of a tone. There is no simple, 1-parameter way of characterizing the generalization in our experiment. One has to deal with structure and to work with a model of generalization over that structure. Our guess is that once structures have been worked out in a particular area, the generalization model will prove to be a natural one for that structure.

In relating our theoretical results to the broader question of syntax learning, we find the notions of "paradigmatic" and "syntagmatic" (e.g., Ervine-Tripp, 1961). Paradigmatic responses are mutually substitutable in a frame. Syntagmatic responses occur next to each other. In response 1, we might say N and I are paradigmatic responses, because either one can occur there. But it is important to realize that, say, I and N in response 3 are not paradigmatic in the same sense. That is, although both can occur in position 3, they are not mutually substitutable, because which one can appear depends on the history of the string. Essentially, paradigmatic responses are responses that fill the same slot in an ordered-state finite automaton. We can generalize this notion by saying that paradigmatic responses fill the same slot in an ordered-state 1-memory store.

I end this section by presenting a summary of our predictions. Figure 5 shows what results lead to what conclusions.

Insert Figure 5 about here

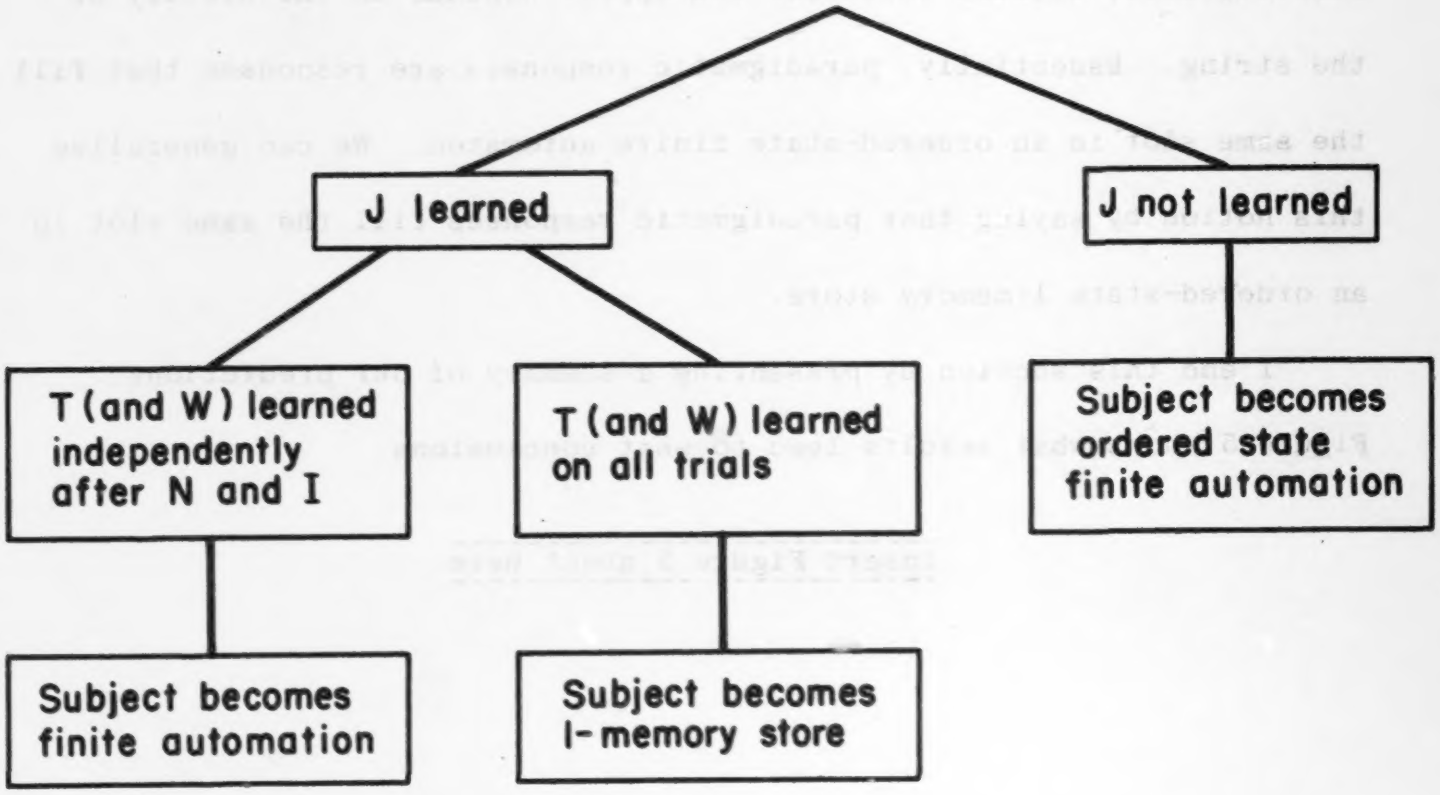
J learned
J not learned
T (and W) learned independently after N and I
T (and W) learned on all trials
Subject becomes ordered state finite automation
Subject becomes finite automation
Subject becomes I-memory store

Fig. 5. Diagram of conclusions to be drawn from various experimental results.

III. Syntactic and Semantic Models

The purpose of this section is to provide another rationale for choosing the kind of system we studied. We discuss a linguistic model which has things to say both about syntax and semantics in natural language and which shows how our miniature system seems to capture some essential properties of that model.

The model was proposed first by Chomsky (1965). We do not discuss the details of how it applies to natural language. Although by presenting the theory in the way we do, we might have a tendency to caricaturize it, the essential ideas should be represented adquately.

Chomsky's proposal is that all natural languages take something like the following form. There is a single, universal syntactic base which, except for lexical entries, is mostly context-free. This base is universal in the sense that all languages have the same base. The context-free base operates first, and then the context-sensitive lexicon rules. The lexical rules (which insert words) of course are specific to each language. At this point we have a collection of phrase-markers. The transformational rules now operate on these phrase-markers, changing the phrase-markers and at the same time the terminal sentences. The transformational rules are specific to each language and are what cause the syntax of different languages to be different. One more assumption (originally proposed by Katz and Postal, 1964) is that transformations do not alter semantics. That is, the meaning of a transformationally derived sentence is not different from the meaning of the sentence it was derived from. Chomsky argues that all semantic interpretation is done on the base. This conception of grammar has recently been challenged by a number of linguists, for example, Lakoff and Ross (1967), who claim that, instead of a base generating syntax with a semantic interpretation, the base should directly generate semantics. Syntactic transformations would be defined to operate on the output of a semantic base. This view is known as "generative semantics" as opposed to "generative syntax." However, only the barest suggestion of formal work has been done from this point of view, for the reason that the problem of semantics representation is almost completely unsolved for natural language. It is not clear how this approach would change the way in which we represent our arithmetic example, and we will ignore it from now on.

We looked for a small domain on which we could experiment that would have as many essential properties of the above system as possible, while holding down the non-crucial complexity as much as possible. This turned us to arithmetic. Arithmetic is taught in almost all, if not all, countries where there is any kind of formal education. It is a simple system which, it turns out, can be cast in a form with just the properties required by this theory. We are talking here about spoken arithmetic, that is, sentences which might be said in a classroom when a teacher is teaching a child arithmetic. It is not true that spoken arithmetic is the same from country to country. The questions are asked in a language, and languages differ. We looked at French, German, and Russian, but in simple arithmetic sentences we did not find much more than different lexical items. That is, there is a function f from V_1 to V_2 where V_1 is the relevant vocabulary of language 1 and V_2 is the relevant vocabulary of language 2 such that if $v_1 \cdots v_n$ is a sentence of language 1, then $f(v_1) \ldots f(v_n)$ is a sentence of language 2, and the two sentences mean the same, that is, their answers are the same. This is not true for these three languages in general, but it is roughly true for the small arithmetic domain we examined. Of course, f is the usual translation function. However, Japanese provided some differences in syntax, and so we settled upon that language.

What is important is that the base of arithmetic is universal across cultures. The part of arithmetic that does not depend on language is universal or almost universal. Specifically, an equation like "2 + 3 = 5" is almost universal in classrooms throughout the world. Even the so-called "Polish" notation in which the above equation would be written as " = + 235" is not used in school classrooms, as far as we know.

Of course, the question, "What does 2 plus 3 equal?" is not universal, but is specific to English. This sentence can be described via a transformation from an underlying sentence such as "x = 2 + 3," which may be an equation in the universal base. The system we propose for arithmetic, in other words, has an underlying context-free base which is roughly universal and generates arithmetic equations. Transformations then operate on this arithmetic base to yield sentences in a specific language. The transformations are specific to each language and thus have to be written for each language. The base, on the other hand, must be constant.

This model can be worked out in practice. We take as our base the rules in Table 1. The notation is standard linguistic notation. Set brackets mean to choose exactly one element inside the brackets. Q_{YN}

Insert Table 1 about here

TABLE 1

Syntactic Base for Arithmetic. The Rules are Ordered and May Apply Any Number of Times.

1. $S \rightarrow \left\{Q_{YN}, Q_{wh}\right\} N = N$

2. $S \rightarrow C(N \begin{Bmatrix} \pm \\ - \\ \cdot \\ / \end{Bmatrix} N)$

3. $N \rightarrow (N \begin{Bmatrix} \pm \\ - \\ \cdot \\ / \end{Bmatrix} N)$

4. $N \rightarrow \{0,1,2,3,4,5,x\}$

represents yes-no questions, Q_{wh} represents what questions, and C represents commands. There is a question, of course, about what a base for arithmetic should contain. We do not claim that there is any particular reason to pick our base over one slightly different. Our point is that the model can be applied, not that we have found the correct solution or even that there is a correct solution. The whole problem of evaluation procedures for grammars could be brought up here, but it would serve no useful purpose.

The base is context-free, as the model requires. Note that it generates many non-true sentences, but it is set up to generate all well-formed sentences, not all true ones. The base generates well-formed sentences for the first 6 (0-5) integers, which are the ones we used in the experiment. It could be modified for any finite number, or a separate system could be written to generate all the integers.

A more difficult task is to write the transformational rules for a given language. One problem is how much to include, since there are many ways of asking arithmetic questions or giving commands in, say, English. We have, fairly arbitrarily, selected some of the more prominent sentences to generate. Once again, the goal has been to demonstrate that the model is applicable, not to yield any kind of complete solution.

Appendix I contains a sketch of the transformational rules for English arithmetic and Japanese arithmetic. Notation is the standard one used in transformational theory (see, for example, Chomsky and Miller, 1963). The sentences generated by the Japanese grammar were obtained from a Japanese informant, who was told to judge sentences on whether they were likely to be heard in an elementary-school class. Others have written a grammar using the same base and same form of transformations for German arithmetic and, partially, for Russian arithmetic.

So far we have seen how the syntactic properties of the linguistic model we are discussing can be incorporated in the arithmetic model. What can we say about semantics? The semantics of the simple arithmetic we are discussing is well understood. The semantic model is the truth model for arithmetic. There are two kinds of base sentences those that contain an x (variable) and those that do not. These two kinds of sentences have different semantic interpretations, analagous in English generally to "what" questions on the one hand and "yes,no" questions on the other. We define the meaning of a base sentence in the following way. Let L(B) be the set of all terminal strings generated by the base. The meaning A is a function from L(B) into the set of subsets of positive and negative integers 2^I, plus the values T and F (for true and false), that is

$$A: L(B) \rightarrow 2^I \cup \{T, F\},$$

meeting the following conditions. Let s be in L(B). Then,

(1) if $s = Q_{YN}E$ for some E, then if E contains an x, $A(s) = \emptyset$, and if E does not contain an x, $A(s) = T$ if E is true, and $A(s) = F$ if E is false.

(2) if $s = Q_{wh}E$ for some E, then if E does not contain an x, $A(s) = \emptyset$. If E contains an x,

$$A(s) = \left\{\begin{array}{l}\text{rational numbers } i \text{ such that } s \text{ is true when } i \text{ is} \\ \text{substituted for } x \text{ in } s.\end{array}\right\}$$

(3) if $s = CE$, then if E contains an x, $A(s) = \emptyset$, and if E does not contain an x,

$A(s)$ = (the rational number y such that $y = E$ is true).

The meaning of certain terminal strings is empty. For example, when the sentence is yes, no and there is a variable in the sentence, we consider the meaning empty because there is no reasonable answer to such a question, unless the value of the variable has been specified. We are not considering such processes here, though in principle we could. It would involve some linguistic processes not well understood, namely, meaning relations across sentences.

We can paraphrase the three conditions above. Assuming that the proper variable condition holds, we see that the meaning of a yes, no sentence is simply its truth value. The meaning of a "what" question is the set of values that make it true, that is, its answers. If there is exactly one x in s, then $A(s)$ will contain exactly one integer. If there is more than one x in s, then $A(s)$ may contain different numbers of elements. For example, $A(Q_{wh} 3 + x = x) = \emptyset$, the empty set, because there is no value of x which makes this sentence true. On the other hand, $A(Q_{wh}\ x + x = 8) = \{4\}$, and $A(Q_{wh} x + 0 = x) = I$, the set of all integers. The meaning of a command is simply the number obtained by carrying out the operations in the sentence.

Now that we have defined the semantics of the set of base sentences, we can define the meaning of any sentence in the language, that is, we extend A to be a function on $T(B)$, the language generated by the base together with the transformations. If $s \in L(B)$, we let the transformational rules apply to s and obtain the sentence $T(s)$. Let $T^{-1}: T(B) \rightarrow L(B)$ be the function such that if $s' \in T(B)$ and $T^{-1}(s') = s$, then $T(s) = s'$. Our statement that T^{-1} is a function requires that T be a one-to-one function, that is, the transofrmational rules may not take more than one base sentence into a given surface sentence (a surface sentence is one on which the transformations have operated). This is the case with our transformations, and for simplicity, we make this one-one assumption here. However, the assumption is not necessary; instead we could have let $T^{-1}:T(B) \rightarrow 2^{L(B)}$. In this case we would have (semantically) ambiguous sentences, as we will soon see.

Now we can define the meaning of any sentence. Let $s \in T(B)$. Then we define $A(s) = A(T^{-1}(s))$. That is, we extend A to a function on $T(B)$ by taking the meaning of a non-base sentence to be the meaning of the base sentence from which it was derived. We have captured here the semantics assumptions of the linguistic model. The meaning is in the base, and transformations do not change meaning. For example, $A(Q_{wh} 2 + 3 = x) = \{5\}$. Applying English transformations to this base sentence yields "What is two plus three?" By our definition, A (what is two plus three) = 5. Returning to a point we made earlier, if T were not one-one and we had defined T^{-1} more generally as we suggested earlier, we could have generalized A, defining it, in essence, as the set of meanings of the sentences which transformationally map into it. Thus semantic ambiguity. A sentence has more than one meaning when it is derivable transformationally from more than one base sentence.

Perhaps we may say a word more about the semantics groups in our experiment. We suggested in Section I that semantics might help syntax learning by restricting the possible structures. In the experimental language J, the sentences have only one answer, and this restricts the possibilities, given the base, of their syntax. For example, it is unlikely that a sentence would contain "ikutsu" twice and a number only once, because only rarely would such a sentence have exactly one answer. A possible model of what a subject is doing when he is trying to learn semantics in our experiment is that he is looking for the base string which transformationally maps into the sentence he is examining. Since he knows the semantics of the base string (we assume this; surely it is true for our subjects' knowledge of arithmetic), if he can find the base string, he will know the semantics of the surface string. Now, since meaning does not change when transformations are applied, any essential meaning-bearing elements in the base sentence will have to be represented somehow in the surface sentence, or else the meaning will change. For example, if the base sentence contains two numerals, then these numerals, perhaps in some transformed form, must appear in the surface sentence. Therefore, practice on semantics might lead the subject to realize that the strings all have two numerals, and this would tell him something about the syntax which would help him in responses three and five. Thus, if an ikutsu has already appeared, then the third word must be a numeral. Similarly, semantic considerations say something about the fifth word. That is, semantics restricts only words three and five. So it is on these responses that the restriction-of-structures model of semantics predicts that subjects will learn faster.

The main point of this section has been to provide a rationale for studying spoken arithmetic. The miniature system we studied seems to capture many of the essential properties of the linguistic model. Perhaps by studying the learning of the miniature system we will increase our understanding of the learning of natural languages.

IV. Experimental Method

Outline of Experiment

Briefly, the experiment had the following form. Part I consisted of pretraining on the four function words, so subjects could learn to recognize the words in sentences and also so they could be trained to respond with the first letter of the word where appropriate in Part III. Part II was paired-associate learning of the six Japanese numerals from 0 to 5. This was necessary so that the semantics group could learn the semantics of the sentences in Part III. As a control, the non-semantics groups also learned the numerals. This part further allowed the subjects to learn the numerals so they could respond N where appropriate in Part III. Part III presented the sentences slowly one word at a time, and the subjects tried to learn which word or words could come next. The sentence was repeated quickly. The semantics group tried to write the answer, and then saw the correct answer. In case gross differences existed between the semantics and non-semantics groups, three non-semantics groups were run to see if we could pin-point the factor causing that difference. None of the non-semantics groups saw or attempted to give the correct answers. One sub-group did nothing while the semantics group wrote and saw the answers. However, if this group did worse than the semantics group on all the responses of the syntax learning, it might be argued that this was due to a lack of practice in general. The semantics group might have spent more time on a task related to and concerning the same sentences as the syntactic task. Therefore a second sub-group was run which, in the time that the semantics group was writing and seeing the answers, had the task of writing down in order the first letters of the sentence they had just heard repeated quickly. This gave them direct practice on the syntax in an attempt to overcome the stated objection. Both sub-groups were told, as was the semantics group, the basic algebraic nature of the sentences. This might make a crucial difference, and might in fact be the effect of semantics. This is, knowing the algebraic nature of the sentences would very likely aid syntax learning. Therefore, a third sub-group was run which was not told the forms of the underlying equations. This group like the first sub-group received no task during the period that the semantics group was answering. We would expect that this group would do worst on syntax learning. Part IV of the experiment presented various sentences, half of them drawn from Part III sentences, and the other half drawn from sentences containing "ikutsu" twice or, in a few instances, sentences ungrammatical in other ways. The subject's task was to answer 1 for grammatical and 0 for ungrammatical. Then the correct (0 or 1) answer appeared.

Speaker. The speaker was a native Japanese graduate student at Stanford University, who had left Japan for the first time two years before the experiment.

Presentation. The entire sequence of material for the experiment was recorded on videotape and shown to the subjects on closed-circuit television. The only things to appear on the screen were the Japanese speaker and, where appropriate, an integer, e.g., "2." Whenever we refer to "the subject heard" or "the subject saw" or "an integer appeared," we mean with respect to the television screen. When we say an integer appeared on the screen, we always mean in numerical form, e.g., "2," not "two" or "ni."

Subjects. Seventy-three subjects were recruited from the Stanford student placement service. Most of them were either students in summer school or students during the regular academic year. The subjects were run in groups of six to thirteen. All subjects run together were run on the same condition, i.e., either they were in the semantics group or the same non-semantics sub-group.

Procedure. The four parts of the experiment were run sequentially, with each subject participating in all four parts. The entire experiment lasted less than an hour and a half. There was no delay between parts except an interval of less than a minute to collect the subjects' response sheets. Instructions for each part were read at the beginning of that part. Questions were answered, and then the television immediately came on with the beginning of the stimuli for that part. Before the Part I instructions, there were brief instructions informing the subject that this was an experiment in language learning.

Part I - Word Pretraining

Materials were the four Japanese words "ikutsu," "wa," "tasu," and "desuka." The words were spoken five times each, one at a time, for a total of 20 words. There were 3 seconds between each word. The subject was given a sheet of paper with 20 spaces and was told to write the first letter of the word (I,W,T or D). (The words had been read to him in the instructions.) There was no feedback on this part.

Part II - Numeral Pretraining

Materials. The first six Japanese numerals.

zero	-	0
ichi	-	1
ni	-	2
san	-	3
shi	-	4
go	-	5

Instructions. The subjects were told the speaker would say a Japanese number and that they were to learn the English translation. They were to write their answers on a provided sheet of paper and to guess if they did not know the correct answers. They were told the correct answer would appear in numerical form after the period in which they were to write the answer, and they were to write the answer before the correct answer appeared.

Procedure. The numerals were spoken in Japanese by the speaker. An item went like this. A Japanese numeral was spoken. During a 3½-second response interval the subject was to write his response. Then the correct answer (translation), an integer in numerical form, appeared on the lower right-hand of the screen for 2 seconds. The next Japanese numeral was spoken. An example of a trial on the numeral 3 is

speaker says "san" -- a 3½-second pause while subject writes down his answer --

"3" appears on screen for 2 seconds -- next item.

There were 10 trials on each of the numerals for a total of 60 items. The numerals were presented in trials with no break between trials. That is, the six numerals were presented randomly, then re-randomized and presented again; this process was repeated to give 10 trials. The only constraint on the randomization was that a numeral could not appear two times in a row, that is, the same numeral could not end one trial and begin the next.

Part III - Sentence Learning

Materials. The sentences used were of the following three forms

ikütsu tasu N wa N desuka,

N tasu ikutsu wa N desuka,

N tasu N wa ikutsu desuka,

where N stands for any Japanese numeral from 0 to 5. (Different N's in the same sentence were not necessarily the same numeral, of course.) A way to interpret these sentences is to translate "ikutsu" as "what," "tasu" as "plus," and "wa" as "equals," so that the first sentence is "What plus N equals N," the second "N plus what equals N," and the third "N plus N equals what?" When we speak of the correct answer to any of these sentences, it was obtained by finding the correct answer to the translated sentence. For example, recalling that "san" = "3" and "go" = "5," in the sentence "san tasu ikutsu was go desuka," we know the correct answer is "2." According to our Japanese speaker, these sentences would be spoken in an elementary-school arithmetic class. Half of the sentences were chosen from the third form shown above (i.e., NTNWID), and the other half was divided between the other two forms. Note that the third form demanded that the subject add to get the correct answer, and the other two forms demanded that he subtract. Thus, by any constant guessing scheme, if the subjects did nothing but add or subtract the two numbers, the semantics group would be correct half of the time.

Altogether 72 sentences were presented. Using the integers 0-5, we had 6 × 6 = 36 sentences of the form NTNWID. Since we did not want any answer greater than 9, we eliminated the sentence with two 5's to give 35 sentences. Then we repeated one sentence to provide 36 sentences for this form. If we look at the form ITNWND, there are only 21 possibilities because to assure a positive answer, the second N has to be greater than or equal to the first N. We picked 18 of these 21 randomly. Exactly the same was true for the other form, NTIWND, and we also selected 18 of 21 sentences randomly for that form. In this way we obtained a total of 36 + 18 + 18 = 72 sentences.

Instructions. The subjects were divided into four groups in this part. We denote them S (semantics), $\bar{S}W$ (non-semantics, writing), $\overline{SW}$ (non-semantics, non-writing), and $\overline{SA}$ (non-semantics, non-algebra). The Part III instructions for Group S are contained in Appendix II. We review them here and also indicate how the instructions for the other groups differed. All groups except $\overline{SA}$ were informed of the basic algebraic nature of the sentences, that is, that they were arithmetic sentences of the given kind and that they asked questions which were related to certain algebraic equations. See Appendix II for the exact statement to subjects. Group $\overline{SA}$ was told only that they would hear Japanese sentences. Of course the subjects had had training on the numeral translations, so they at least knew there were numerals in the sentences.

The subjects in all groups were instructed that they were to hear actual Japanese sentences, and that their job would be to learn the patterns of these sentences. They were to hear each sentence word by word, and they were to predict what words could come next after they heard each word. They were told that in each pattern at most two words (perhaps one) could be correct. They were told to write their answers in the boxes provided. The boxes had room for two responses, separated by a comma. If they thought only one word could be correct, they were to predict that word. If they thought either of two words could be correct, they were to predict both of the words. (Note that at any

2.

point only one word would actually be said in a sentence, but the patterns were such that sometimes another word *could* have been said.) The subjects made their predictions by writing the first letter of the word in the appropriate box on the sheet provided if they wanted to predict ikutsu, wa, tasu, or desuka. If they wanted to predict a numeral, they did not write the first letter of the number, but wrote N. To repeat, the subjects were told that they could write either one or two of the letters I, W, T, D or N at each point.

At this point instructions among groups differed. First, the semantics group was told that after they finished the above procedure for a sentence, they would hear exactly the same sentence repeated, but this time more quickly, at a fairly natural rate. After they heard the sentence repeated, they were to write the answer to that sentence, a digit from 0 to 9, in the space provided. If they did not know the answer, they were to guess. In a few seconds the correct answer would appear on the screen, and they were to try to learn so that they would be correct.

Groups $\overline{\mathbf{SW}}$ and $\overline{\mathbf{SA}}$ were told that the sentence was repeated to help them learn it. They had no other task before the next sentence started. Group $\overline{\mathbf{S}}\mathbf{W}$ was told the same thing, but had the task of writing the first letters of the sentence they had just heard in spaces provided for them, with the digits not N, actually being written.

Procedure. The number of subjects in each group is given in Table 2. The subjects were assigned randomly to groups to the extent possible,

Insert Table 2 about here

TABLE 2

Number of Subjects in Each Group.

			Total		
$\bar{S}W$	$\overline{SW}$	$\overline{SA}$	$\bar{S}$	S	Total
13	13	13	39	34	73

given the times that they could appear for the experiment, which was run in groups of 6 to 13 subjects. Each group was provided with paper marked for the responses they were instructed to make. For example, none of the $\overline{S}$ groups had room for numerical answers to the sentences. The spaces for the predictions for the next possible words contained, for each position, a box with a comma in the middle so that subjects could put in either one or two responses.

A trial started by a tone sounding. The subjects were given 4 seconds to make their predictions of the first word of the sentence. Then the first word of the sentence appeared (i.e., it was said by the speaker on the screen). Again the subjects were given a 4-second pause to write their predictions for the second word. The second word was said, and so on, until the end of the sentence. After the sentence was finished there was a 2-second pause, and then the sentence was repeated by the speaker, but this time at a normal rate of speech.

For the semantics group (S) there was now a second pause of 4 seconds, during which the subjects wrote the answer (a digit from 0 to 9) to the sentence they had just heard. Then the answer appeared on the lower right of the screen for 2 seconds. After a 1-second pause the tone sounded to begin the next trial. A diagram for the sequence of events for the example "san tasu ikutsu wa go desuka" appears in Figure 6.

Insert Figure 6 about here

Up to the point after the sentence was repeated, the procedure was the same for the non-semantics groups as for the semantics group. However, the answer did not appear on the screen for the non-semantics group, and the subjects did not have the answering task. Exactly the

Television	tone sounds		san		tasu		ikutsu		wa	
Comment	trial starts		S1		S2		S3		S4	
Subject		N,I		T		N,I		W		N
Comment		R1		R2		R3		R4		R5
Time		4s		4s		4s		4s		4s

Television	go		desuka		san tasu ikutsu wa go desuka		2		tone sounds
Comment	S5		S6	pause	sentence repeated faster		correct answer	pause	next trial starts
Subject		D				2			
Comment		R6				number response			
Time		4s		2s	approx. 2 to 3 sec.	4s	2s	1s	

Figure 6. Diagram for the sequence of events for one trial for Part III, Group S (semantics) on the sentence "san tasu ikutsu wa go desuka." The responses given for the subject are those he would give if he were correct. In the time row, "s" means "seconds."

same videotape was used for the non-semantics groups as for the semantics group, but for the non-semantics group the answer was covered, so that there would be no difference in presentation between the two groups except for the appearance of the answer. Thus after the sentence was repeated, for the non-semantics groups ($\overline{S}$ groups), there was a pause of 4 seconds (as for the semantics group), plus 2 seconds (the covered answer was on) plus 1 second (as in the second pause for the semantics group) for a total pause of 7 seconds. During this time groups $\overline{SW}$ and $\overline{SA}$ had no task. Group $\overline{S}W$ had to write the sequence of the first letters of the words in the sentence they had just heard repeated. For example, if they heard, "san tasu ikutsu wa go desuka" they should have written "3TIW5D."

The 72 randomized sentences were presented in this fashion. All the subjects had the same order of presentation of sentences; indeed, the tape was the same for all subjects.

Part IV - Grammaticality Learning

Materials. Fifty sentences were used. Twenty-four of them were "grammatical" (G) and 26 "ungrammatical" (U). (There were supposed to be 25 of each, but a mistake was made in the recording.) The 24 G sentences were chosen randomly from the kinds of sentences used in Part III; 8 of each form were chosen. Of the 26 U sentences, 22 were selected from Part III, grammatical form, substituting "ikutsu" for one of the numbers; a typical example might be "ikutsu tasu san wa ikutsu desuka." These 22 were about equally divided (7, 7 and 8) among the three kinds of sentences whose original grammatical sentence was one of the three kinds of Part III sentences. These kinds of ungrammatical sentences were chosen, because if a subject became the kind of ordered-state automaton we discussed in Section II (Figure 3), he would consider these sentences grammatical.

The other 4 U sentences were chosen by permuting two words in a grammatical sentence. The sentences were

ikutsu desuka 1 wa 3 tasu,
0 5 tasu wa ikutsu desuka,
2 ikutsu tasu wa 3 desuka,
ikutsu wa 2 tasu 4 desuka.

The 50 sentences were randomized; the only restraint was to present the 4 special U sentences at least 8 sentences apart.

Instructions. The subjects were told that in this part they would use some of the knowledge they learned in Part III. They were told they would hear Japanese sentences, and "your job is to determine if these sentences are exactly like the sentences you heard before in Part III. That is, could this sentence you hear have been one you heard before? If yes, write a 1 in the box. If no, write a 0." They were told the correct answer would then appear on the screen.

Procedure. The 50 sentences were presented randomly as described above. The subjects were given sheets of paper to write their answers on. A trial went like this. A sentence was spoken. There was a 3½-second interval during which the subjects were to write their answers (1 or 0). Then the correct answer (1 or 0) appeared on the lower right of the screen for 2 seconds. The next sentence was read and the cycle repeated.

V. Experimental Results

Part I - Word Pretraining

On Part I, out of a total of 1,460 responses (20 responses × 73 subjects), there were only 11 errors. Clearly the task was extremely easy, and subjects had no trouble discriminating the words.

Part II - Numeral Pretraining

The learning curve for Part II appears in Figure 7.

Insert Figure 7 about here

Clearly an asymptote of no errors has been approached. On the last 3 trials there is a mean number of 2.5 errors per trial out of a total of 73 possible.

Part III - Sentence Learning

Syntax Responses. We call the responses the subjects made in predicting the next possible words their *syntax* responses, as opposed to the *semantics* responses, which were the number answers for the semantics group. The form of the data is the following. There were 72 trials for each subject, and for each trial six words were presented, which we call the stimuli, and signified, in order of presentation, S1,..., S6. A subject made a "response" which is *blank* or a 1- or 2-element subset of the letters N,I,T,W.D. In fact, all the responses were of this form, and there were no other letters used by subjects. Further, no subsets of size greater than 2 were used. (The form of the response sheets helped to insure this.) For simplicity we will not use set notation, but write, for a response, e.g., R = I,N instead of $R = \{I,N\}$. The six responses are labeled R1, R2,..., R6, in the order they were made on a trial.

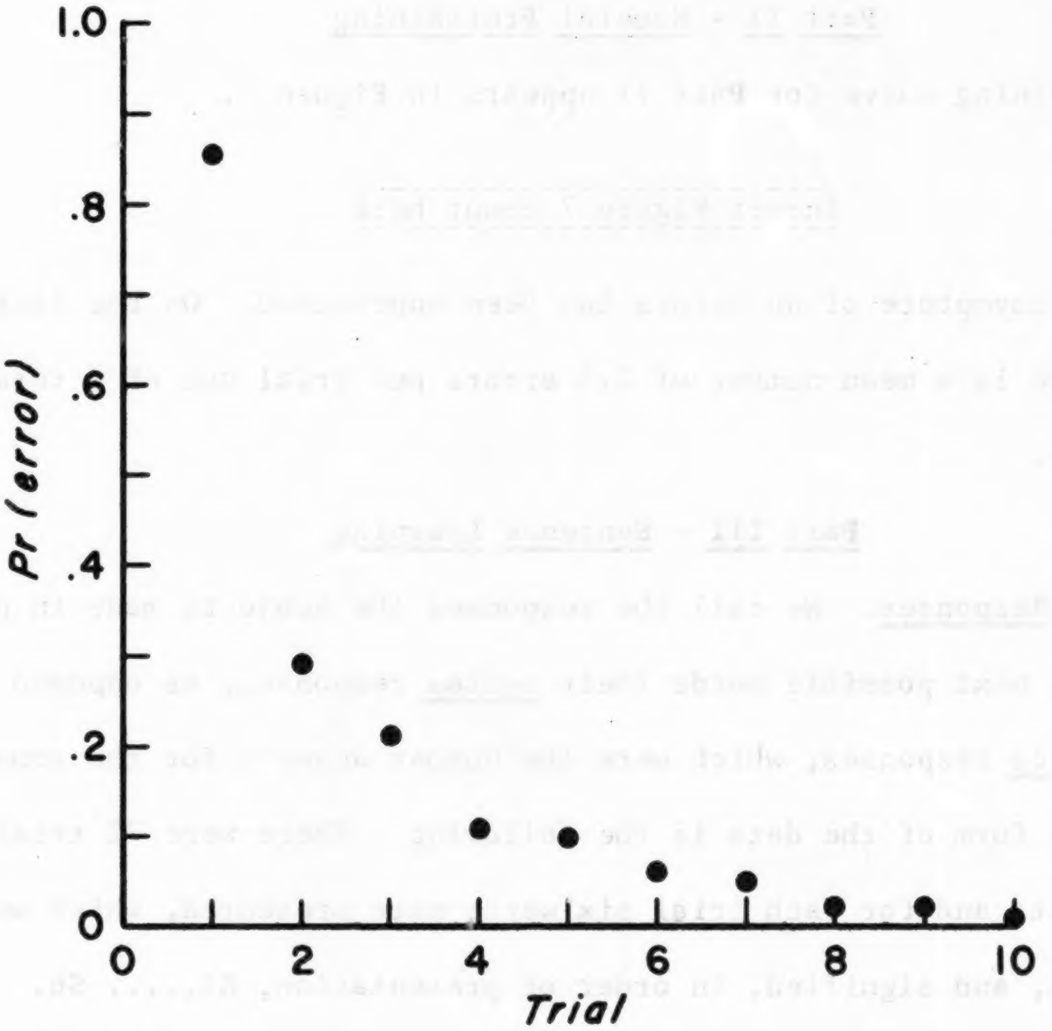

Fig. 7. Learning curve for paired-associate numbers.

Recall that Ri precedes Si. When we count responses of various kinds, if the response contained two elements, we ignore, as the set notation implies, the order of the subject's response and count both orders together, e.g., R3 = I,N means either the third output of the subject on the trial was I,N or it was N,I.

Let us first look at whether R3 was learned. The relevant figures are in Table 3. The first row shows the number of subjects in each group.

Insert Table 3 about here

Before we determine whether a subject learned we have to decide if he followed the instructions. Some subjects never put two responses in the same box on any of the 72 trials for any of the six responses, that is, they never made two predictions for the next word. These subjects, of course, could never have learned by our definition. It seemed reasonable to decide that these subjects had not understood the instructions and did not realize that they could put two responses in the same box. Therefore, these subjects were not included in consideration of whether subjects learned. Out of 73 subjects, 14 fell into this category, leaving 59 subjects who followed the instructions. These figures are broken down for the S and $\overline{S}$ subgroups in Table 2. $\overline{S}$ indicates all 3 $\overline{S}$ subgroups combined.

We set the following criterion for learning R3. When S1 = I, then R3 = N is a correct response. If S1 = N, then R3 = N,I is a correct response. If, somewhere in a subject's 72-trial response protocol there is a sequence of 6 or more consecutive correct R3 responses, including responses to at least 2 sentences of each kind, we say the subject learned R3. Most of the responses in this sequence generally will be N,I since

TABLE 3

Number of Subjects on Part III in Various Categories.

	$\overline{S}W$	$\overline{SW}$	$\overline{SA}$	Total $\overline{S}$	S	Total
Total Subjects	13	13	13	39	34	73
Subjects who did not use 2 responses	2	1	4	7	7	14
Subjects who followed directions	11	12	9	32	27	59
Subjects who learned R3	11	8	7	26	24	50
Proportion of subjects following directions who learned R3	1.00	.67	.78	.81	.89	.85

more sentences begin with N than with I, but the criterion requires that at least two of them be N and that these be in sentences starting with I. This requirement is made so that a subject cannot be considered to have learned simply by always saying N,I regardless of S1.

By this criterion, Table 3 shows that 6 subjects in Groups $\overline{S}$ and 3 subjects in Group S did not learn. In other words, 50 of the 59 subjects (85 percent) who understood the instructions learned. Eighty-nine percent of the S subjects and 81 percent of the $\overline{S}$ subjects who understood the instructions learned. There is no significant difference between the S and $\overline{S}$ groups ($\chi^2 = 1.56$, 1 df, $p > .20$). There is also no significant difference from chance on this statistic between the three $\overline{S}$ sub-groups ($\chi^2 = 1.07$, 2 df, $p > .50$). Of course, there are relatively few subjects in each group, when we consider these subgroups. Also, the fact that there is no difference between groups on this statistic does not mean that there is no difference in learning among the groups. The learning rates could still differ. We have provided evidence that most subjects learned R3, and that groups did not differ on how many subjects learned R3.

Of the 9 subjects who followed directions but did not learn, inspection of the response protocols showed that by the end of the 72 trials, 3 of the subjects consistently responded N,I for R3, independent of S1. The other six subjects did not reveal any particular pattern. It seems possible that the three subjects responding N,I were at asymptote and would not change their responses if more trials were added. Since the other 6 subjects were not caught in a pattern, they might have learned the correct responses if more trials were added. In fact, some of these subjects almost met the criterion of learning when the trials ran out.

Figure 8 shows the learning curve for the 50 subjects who met criterion. The asymptote is almost 0, except for an occasional, possibly

Insert Figure 8 about here

"accidental" error. It seems reasonable to conclude that these errors are "accidental," i.e., that the subject learned, but for some reason such as lack of attention due to boredom, did not make the correct response. (A number of subjects complained that the experimental task was too easy.) The learning curve merely shows in another way that these 50 subjects learned the correct response for R3.

R5 enters into our theoretical predictions in the same way as R3, so we turn to it now. We say that a response is correct if, when S1 = I and S3 = N or when S1 = N and S3 = I, the response is R5 = N, or when S1 = N and S3 = N, the response is R5 = 1. The criterion was the same as for R3. A subject learned R5 if he had a sequence of at least 6 consecutive correct responses which included at least 2 N responses and at least 2 I responses. By this criterion, none of the subjects who did not learn R3 learned R5. Of the 50 subjects who learned R3, all but 2 learned R5. Once again, we see that most of the subjects who followed directions learned by this criterion. In the case of R5, 48 of 59 subjects learned.

From now on we will consider the data of only those 50 subjects who learned R3, because we do not know how to interpret the data of the subjects who understood the instructions but did not learn. This involves considering two subjects who did not learn R5, but for simplicity, and so that we could use the same subjects on all tests, we included all 50 subjects even when considering R5. In Figure 9 appears the learning

Insert Figure 9 about here

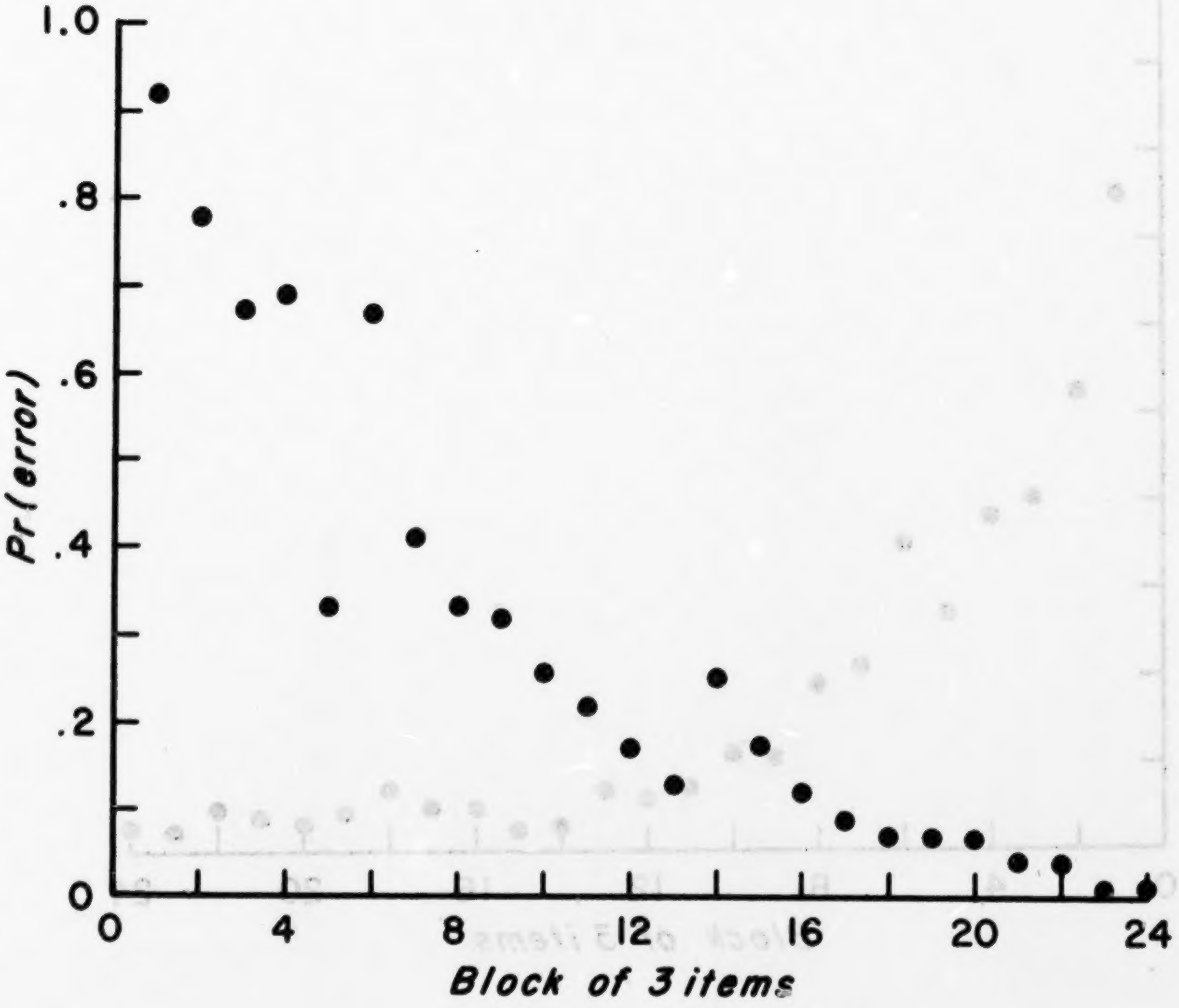

Fig. 8. Learning curve for R3. Some of the roughness in the curve is due to different kinds of items.

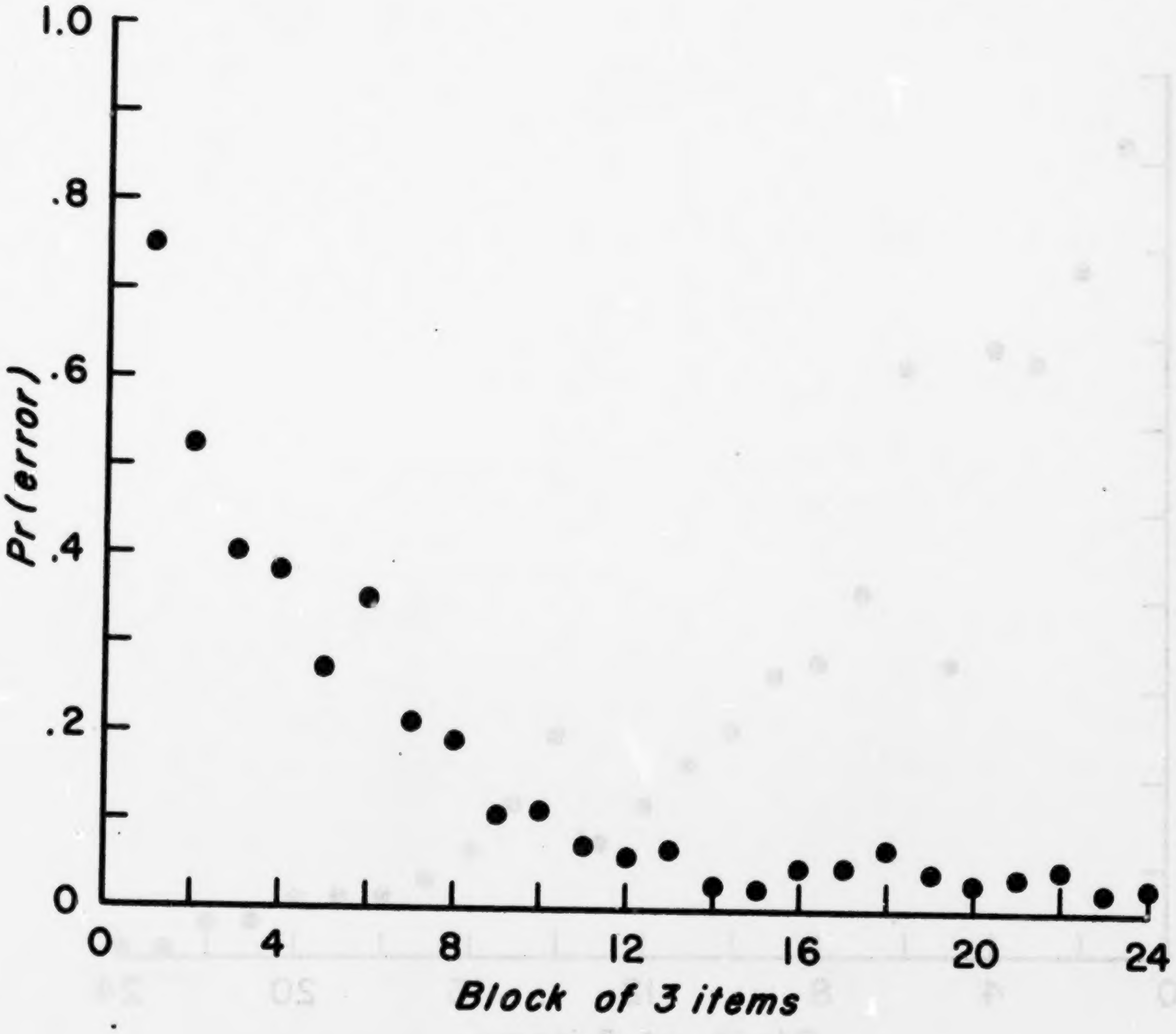

Fig. 9. Learning curve for R5. The curve contains different kinds of items.

curve for R5 for the 50 subjects. Once again the curve shows that subjects learned. It is important to realize, when comparing this curve with the curve for R3 (Figure 8) that although both curves plot the proportion of "correct" responses, the correct responses differ for the two graphs, and in fact, differ from trial to trial within each graph. For R3, the correct responses are N or N,I and for R5, the correct responses are N or I. The fact that the correct response set differs for R3 and R5 reduces even more the probability of subjects giving a correct sequence by chance. That is, we cannot compute the probability of subjects giving a correct sequence by chance as if, for example, in R3, there is a probability p that the response is N and a probability 1-p that the response is N,I, and, for R5, there is a probability q that the response is N and a probability 1-q that the response is I. We cannot simply do this because this does not account for the subject's learning the response set in the first place. S3 was always either N or I as was S5, so there was no way for the subjects to learn the response sets strictly from a consideration of what S3 or S5 could be.

A summary of these results is that, in general, subjects learned both R3 and R5. Also, there was little tendency for subjects, at asymptote, to respond N, I independently of the preceding sequence of words.

In Table 4 we list the mean trial of last error, L, for the six responses for each group. As mentioned earlier, there are 50 subjects

Insert Table 4 about here

in the table. For responses R3 and R5, the trial of last error for each subject is determined by the same method as described earlier for the learning criterion; that is, the trial of last error is the trial before

TABLE 4

Mean Trial of Last Error, L, by Response and Subject Group.

	$\bar{S}W$	$\overline{SW}$	$\overline{SA}$	Total $\bar{S}$	S	Total S,$\bar{S}$
R1	16.5	27.8	7.4	17.5	19.0	18.3
R2	6.7	8.3	6.6	7.2	8.4	7.8
R3	29.6	33.3	28.7	30.5	28.0	29.3
R4	8.1	15.1	5.7	9.6	10.5	10.2
R5	21.4	22.0	25.0	22.5	14.1	18.5
R6	3.7	3.4	5.4	4.1	4.6	4.3
R1,R3,R5	22.4	27.7	20.4	23.5	20.4	22.3
R2,R4,R6	6.2	8.9	5.9	7.0	7.8	7.4
Grand Mean	14.3	18.3	13.1	15.3	14.1	14.9

the occurrence of the first run of at least six correct responses which include at least two of each kind of correct response. For R1, R2, R4, and R6, for each of which there is only one correct response, L is simply the trial before the start of the first run of six or more correct responses.

Table 2 shows clearly that responses R2, R4 and R6 (the "even" responses) were learned more quickly than were R1, R3 or R5 (the "odd responses"). The mean of L for the odd responses for Group $\overline{S}$ (23.5) is more than 3 times as great as the mean for the even responses (7.0). For Group S, the ratio is almost as great (20.4 to 7.8). In fact, if we look at the means for each response we see that none of the 3 even responses has a mean L value as great as any of the 3 odd responses. This last statement holds also within each sub-group of $\overline{S}$. For any group, there are 6! possible ways of ordering the 6 responses with respect to L. Thirty-six of these yield orders compatible with the above statement; that is, the odd values are all greater than the even values. Thus, if we assume the orders were chosen uniformly, the probability of obtaining an ordering compatible with the statement is $36/6! = .05$. Since there were four independent groups (three subgroups of $\overline{S}$ plus S), the probability of obtaining our results by chance is $(.05)^4 < 10^{-5}$.

Inspection of the distributions of L show that there are a few fairly high values. To make sure the results we report for L are not unduly influenced by these high values, we also calculated medians for all the values. The medians are shown in Table 5. The pattern of the results is the same as for the means shown in Table 3. Therefore, we do

Insert Table 5 about here

not discuss these values, but instead concentrate on the means.

Table 6 shows the mean number of total errors, T, for each group.

Insert Table 6 about here

This statistic behaves almost exactly like L with respect to the questions we have been considering. Subjects made many more errors on the odd responses than on the even.

In computing the trial of last error, L, for R3 and R5, we demanded a criterion of 6 in a row correct, including at least two of each kind of trial. This may have caused L to be slightly higher for R3 and R5 than for the other responses. But this is a very small effect. We recomputed L for R3 and R5, relaxing the requirement of two of each kind of trial, and found that the pattern of results did not change. This criticism does not apply to the computation of the statistic T.

Are the mean trials of last error smaller for $\overline{S}$ than for S? Generally, no, as may be seen from Table 4. Table 7 shows values of student's t for the difference between means for the six responses.

Insert Table 7 about here

For 50 subjects, the only significant value is for R5 ($p < .05$). In fact, the other t values are much smaller than R5's. The only other response for which the mean value of L is greater for $\overline{S}$ than for S is R3. These results show that, in general, the S group did not learn faster than the $\overline{S}$ groups. Figure 10 shows the learning curves separately for the S (24 subjects) and $\overline{S}$ (26 subjects) groups for the six responses.

Insert Figure 10 about here

TABLE 5

Median Trial of Last Error, L.

				Total	
	$\bar{S}W$	$\overline{SW}$	$\overline{SA}$	$\bar{S}$	S
R1	10.0	23.5	6.0	10.5	16.0
R2	4.0	7.0	5.0	5.5	6.0
R3	21.0	32.0	23.0	28.5	26.5
R4	5.0	15.5	5.0	6.5	8.5
R5	21.0	17.5	24.0	21.5	12.0
R6	3.0	1.5	5.0	2.5	3.0

TABLE 6

Mean Number of Total Errors, T.

	$\overline{S}W$	$\overline{SW}$	$\overline{SA}$	Total $\overline{S}$	S	Total S,$\overline{S}$
R1	15.4	27.9	7.0	17.0	18.5	17.7
R2	6.5	9.0	9.1	8.0	7.3	7.6
R3	20.6	23.9	18.0	21.0	20.3	20.6
R4	7.5	9.6	5.4	7.6	8.8	8.2
R5	12.5	11.3	18.3	13.7	9.7	11.8
R6	4.5	2.5	6.6	4.5	3.9	4.2
R1,R3,R5	16.2	21.0	14.4	17.2	16.2	16.7
R2,R4,R6	6.2	7.0	7.0	6.7	6.6	6.7
Grand Mean	11.2	14.0	10.7	11.9	11.4	11.7

TABLE 7

Values of t for the Difference Between Mean Trial of Last Error, L, of the S and $\bar{S}$ Groups on R3.

Response	R1	R2	R3	R4	R5	R6
t	0.26	0.73	-0.52	0.38	-2.16	0.46

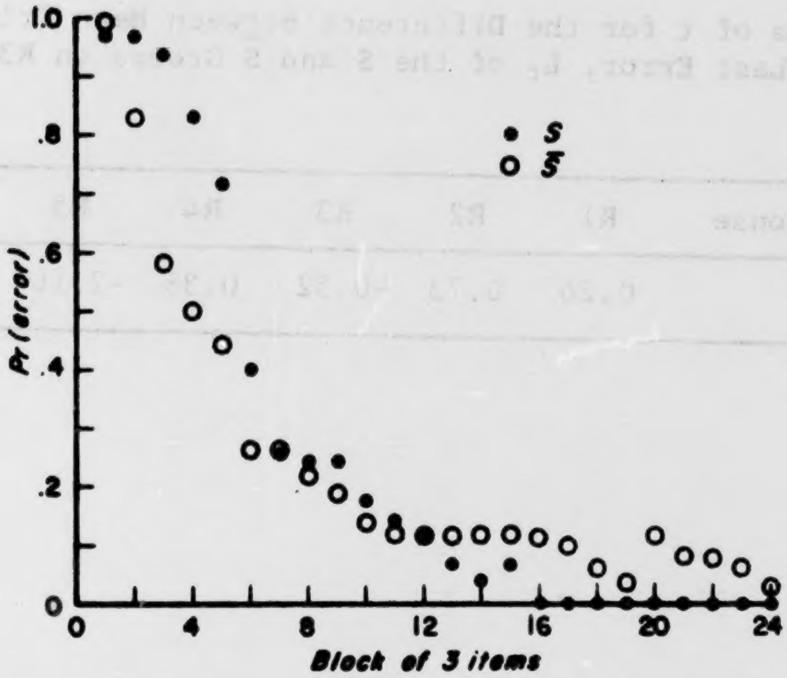

Fig. 10a. Learning curves for R1.

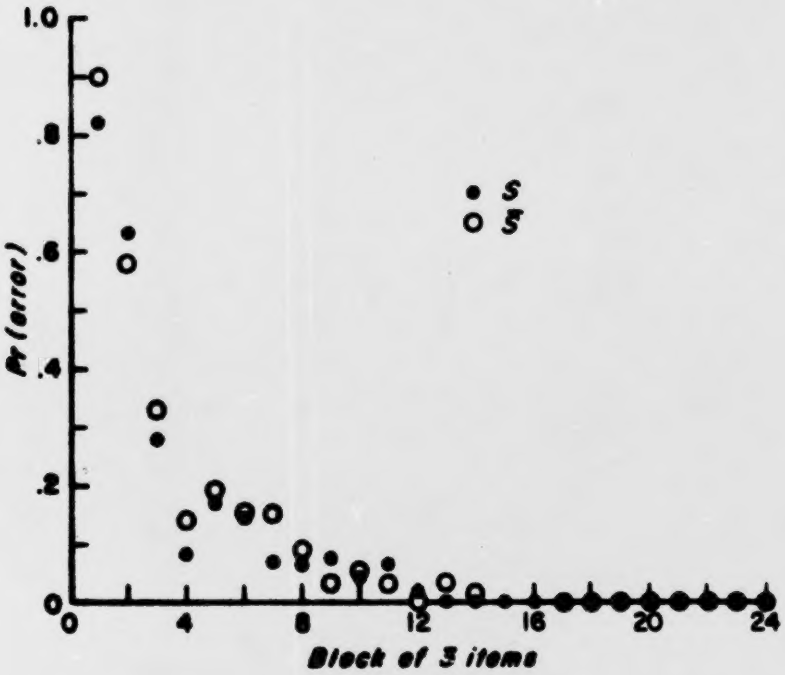

Fig. 10b. Learning curves for R2.

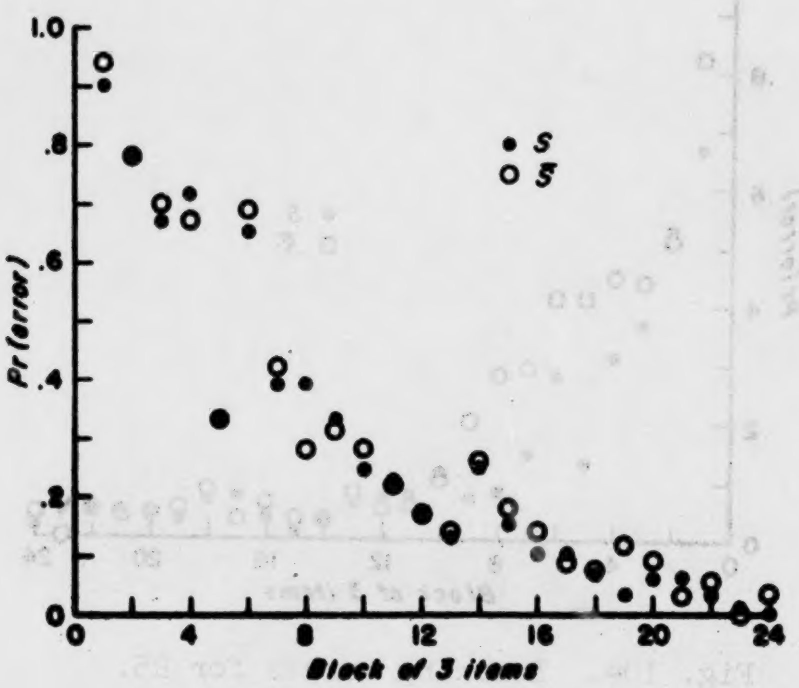

Fig. 10c. Learning curves for R3.

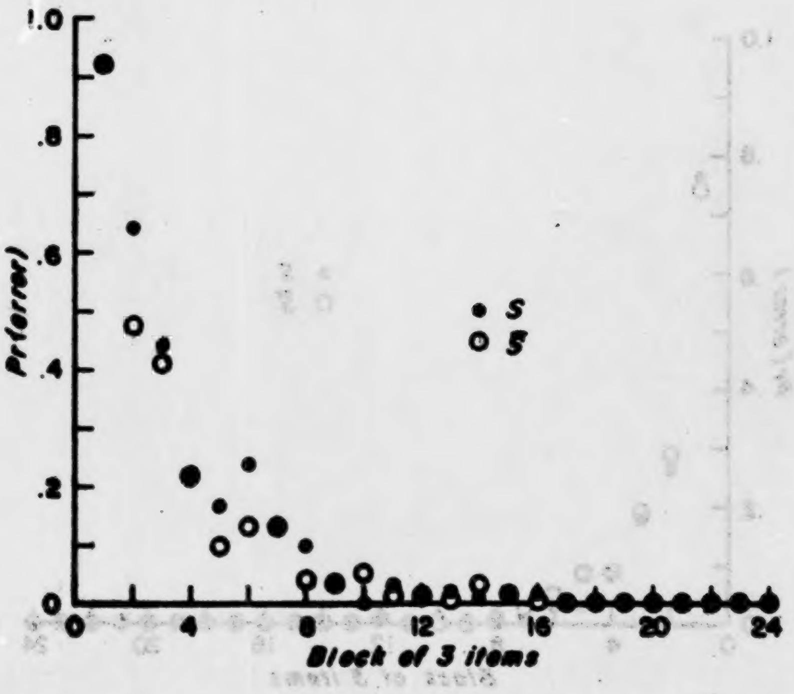

Fig. 10d. Learning curves for R4.

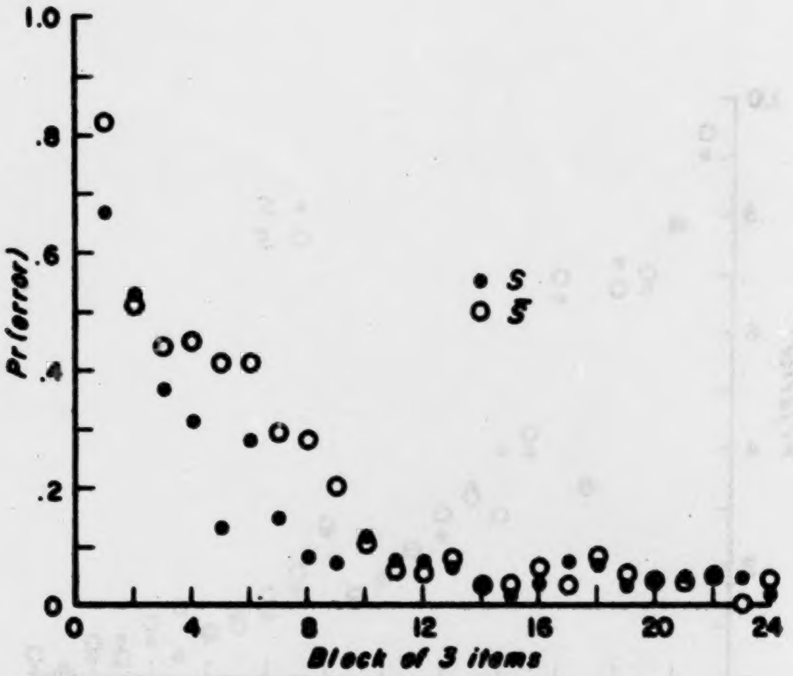

Fig. 10e. Learning curves for R5.

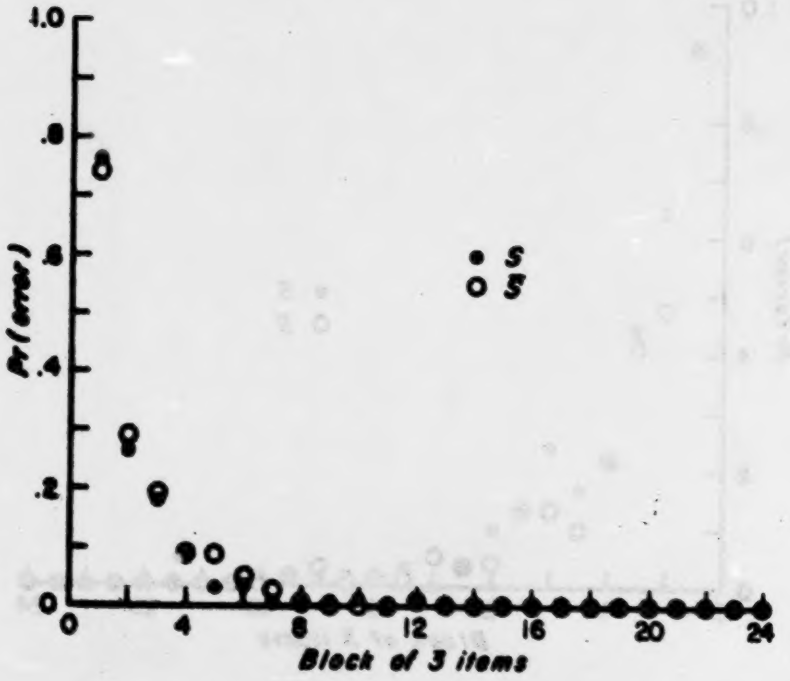

Fig. 10f. Learning curves for R6.

These curves as well as the mean number of total error (Table 6) fit the same pattern of results with respect to the differences between S and $\overline{S}$.

Remembering that two of the subjects did not learn R5, it occurred to us that this may have somehow influenced the results concerning the difference between the S and $\overline{S}$ groups on R5. We included these two subjects in the data, and took as their trial of last error, since they did not meet criterion, the actual last trial of the 72 on which an error occurred. It turned out that this value was 70 for both subjects, and both subjects were in Group $\overline{S}$. Although the subjects in the table were chosen statistically so as not to favor Group S (they were chosen on the basis of whether they had learned R3), it might be argued that accidentally subjects who had not met criterion on R5 were selected for $\overline{S}$ and this pushed up the mean value of L for R5. Therefore, we did a new calculation of L for R5 for Group $\overline{S}$, discarding these two subjects, and calculating the mean L for the 24 remaining subjects. The new value was 18.5 for L, which, compared to the 14.1 for Group S, still yields the largest discrepancy between L for S and $\overline{S}$ of any response. Therefore, even if one accepts this argument, Group S did better on R5 than Group $\overline{S}$ did.

As explained earlier, we ran S in 3 different subgroups under different conditions, so that in case S learned faster than $\overline{S}$, we could see if the difference could be explained by a particular factor. If we look at the mean L value over all responses, group $\overline{SW}$ had the highest value (18.3) and group $\overline{SA}$ had the lowest value (13.1). However, as we stated in the previous paragraph, the only significant difference between $\overline{S}$ and S was on R5, and on this response the mean values of

L for the 3 $\overline{S}$ subgroups are about equal, and all are much greater than for Group S. Since there is no explainable difference between S and $\overline{S}$ by these $\overline{S}$ control groups, we do not consider these subgroups, but lump the data and consider only the one $\overline{S}$ group. One short point can be made about Group $\overline{SA}$ however. Since this group did not even know the algebraic character of the sentences, we had expected them to do worst on the syntax responses; but, in fact, their score was best. However, note that on R5 their mean trial of last error is higher than for the other two subgroups.

In analyzing the difference between the S and $\overline{S}$ groups, we work on the assumption, of course, that because the groups were chosen randomly, there was no difference between the groups except for the different treatment in the experiment. However, we have some direct evidence. Part II of the experiment was conducted before there had been any different treatment for the different groups. By looking at differences in the learning of Part II, we could see if there was any evidence of differences between the groups not related to experimental treatment. In Table 8 we show the mean number of total errors for Part II for Groups S and $\overline{S}$, for subjects who learned R3 and for subjects who did not learn R3 (including those who did not follow instructions).

Insert Table 8 about here

The results are summarized by saying that subjects who did not learn R3 made more errors on the number learning, and subjects in Group $\overline{S}$ made more errors than subjects in Group S. (Between learned and did not learn, $t = 1.96$, $.05 < p < .1$; between S and $\overline{S}$, $t = 1.17$, $p > .1$).

TABLE 8

Mean Number of Errors on Number Paired-Associates (Part II) for Groups S and $\overline{S}$ for Subjects Who Learned R3 or Did Not Learn R3.

	S	$\overline{S}$
Learned R3	8.1	10.3
Did Not Learn R3	11.7	13.3

These results suggest that subjects who did not learn R3 were poorer learners in general (whether for motivational or other reasons we do not know), and that subjects in Group $\overline{S}$ probably were slightly poorer learners than subjects in Group S. The fact that $\overline{S}$ subjects did better on four of six responses in Part III together with this last fact once again suggests that semantics does not have a general improving effect on syntax learning.

We have seen that R2, R4 and R6 were learned faster than R1, R3 and R5. This finding agrees with the prediction made from the 1-memory store model. It is not the case however, that the only difference between the even and odd responses is the one that led to our prediction. The correct response for R1 contains two components (I,N), and one of the correct responses for R3 also has two components (I,N). But the even responses have only one correct response (T,W or D). There may be something which caused subjects to be less ready to respond with two letters than with one. R5, however, did not meet this difficulty. Both correct responses are only a single letter (I or N), and R5 was learned more slowly than any of the even responses. This built-in control thus helped us decide that the difference between the even and odd responses was due to the even responses being learned in such a way that trials with different pasts contributed to learning. In other words, the Equations (2) in Section II are more correct than the Equations (1).

However, there is an even more direct way to test this, as we showed in Section II, and that is to look at whether, say, T was learned independently on trials with different histories. Figure 11 shows the learning curves for Groups $\overline{S}$ and S for R2, R4 and R6 for the first 10

Insert Figure 11 about here

trials. (After 10 trials on these responses there were relatively few errors.) The abscissa is trial number, and the ordinate is proportion of errors. The trials on which S1 = I (that is, the first word presented is I), are plotted by x's. These are trial numbers 1,4,7. The other trials are plotted by dots. Now, if the responses for the two kinds of trials were learned independently, the learning curve would not be a monotonically decreasing curve. Rather, points 4 and 7 would jump way up. In fact, if we assume that the learning rates were equal for the two kinds of trials, the trial-4 point would jump up to the trial-3 point, and the trial-7 point would jump up to the trial-5 point (assuming no interference). On the other hand, if all the trials (i.e., both kinds) count equally toward the learning of the response (i.e., if we assume that all the trials form a sequence of learning trials on the same response), then we should obtain a monotonically decreasing learning curve of the usual kind, with trials 1,4 and 7 falling into place. The curves plotted in Figure 11 show that this latter result is the case. The S1 = I trials appear as they would if the ten trials were a learning sequence on one response.

As a comparison, in Figure 12 the learning curves for the first 10 trials for R3 and R5 are plotted. For R3, the x's are trials on which

Insert Figure 12 about here

S1 = I and dots are trials on which S1 = N. It is clear that the curve here is not monotonic, rather the x points are much lower than the dots. In the R5 curve, the x's are trials on which either S1 = I or S2 = I, and

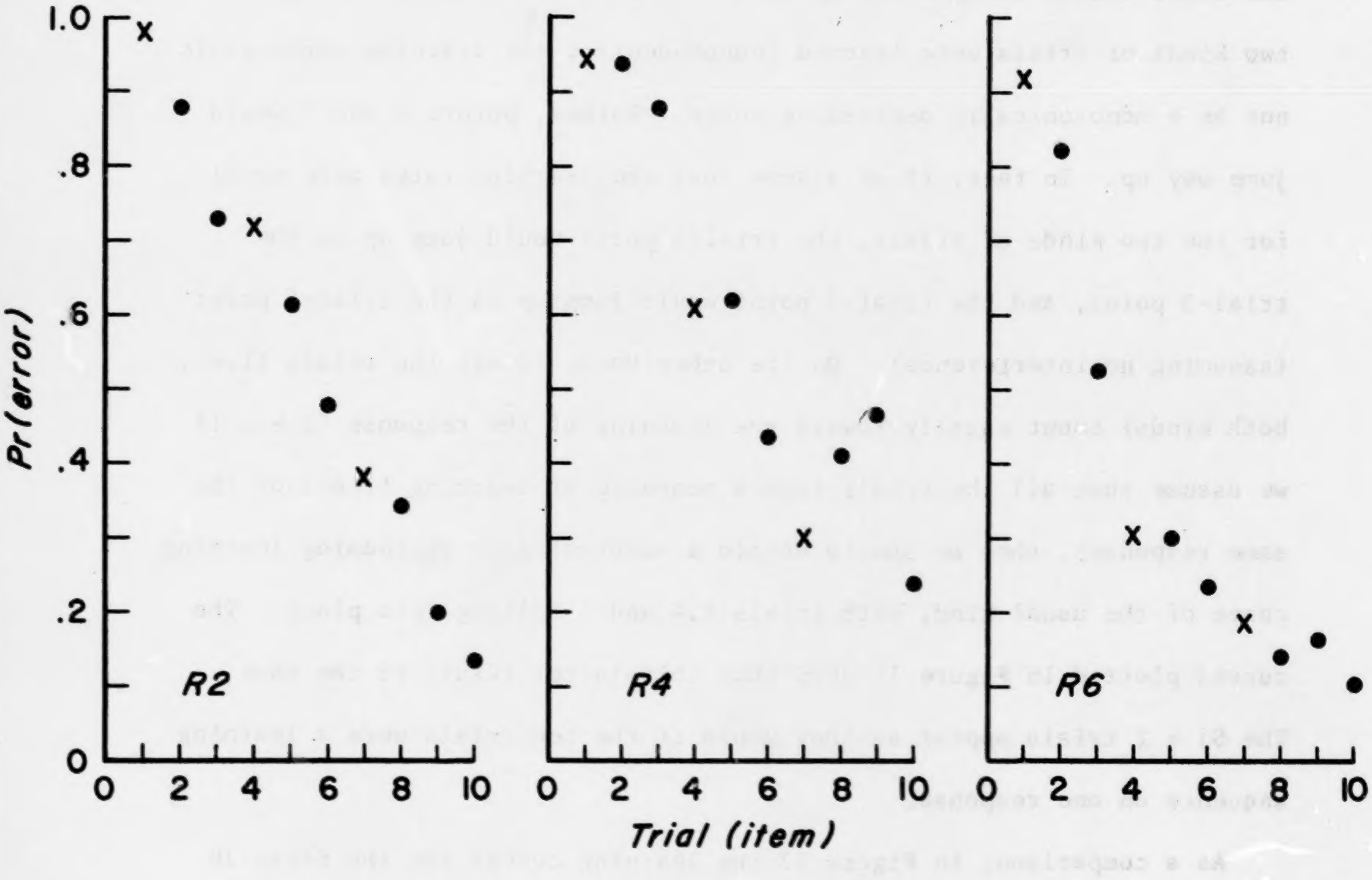

Fig. 11. Learning curves for first 10 trials (items) for R2, R4 and R6. x's are trials on which S1 = I. Dots are trials on which S1 = N.

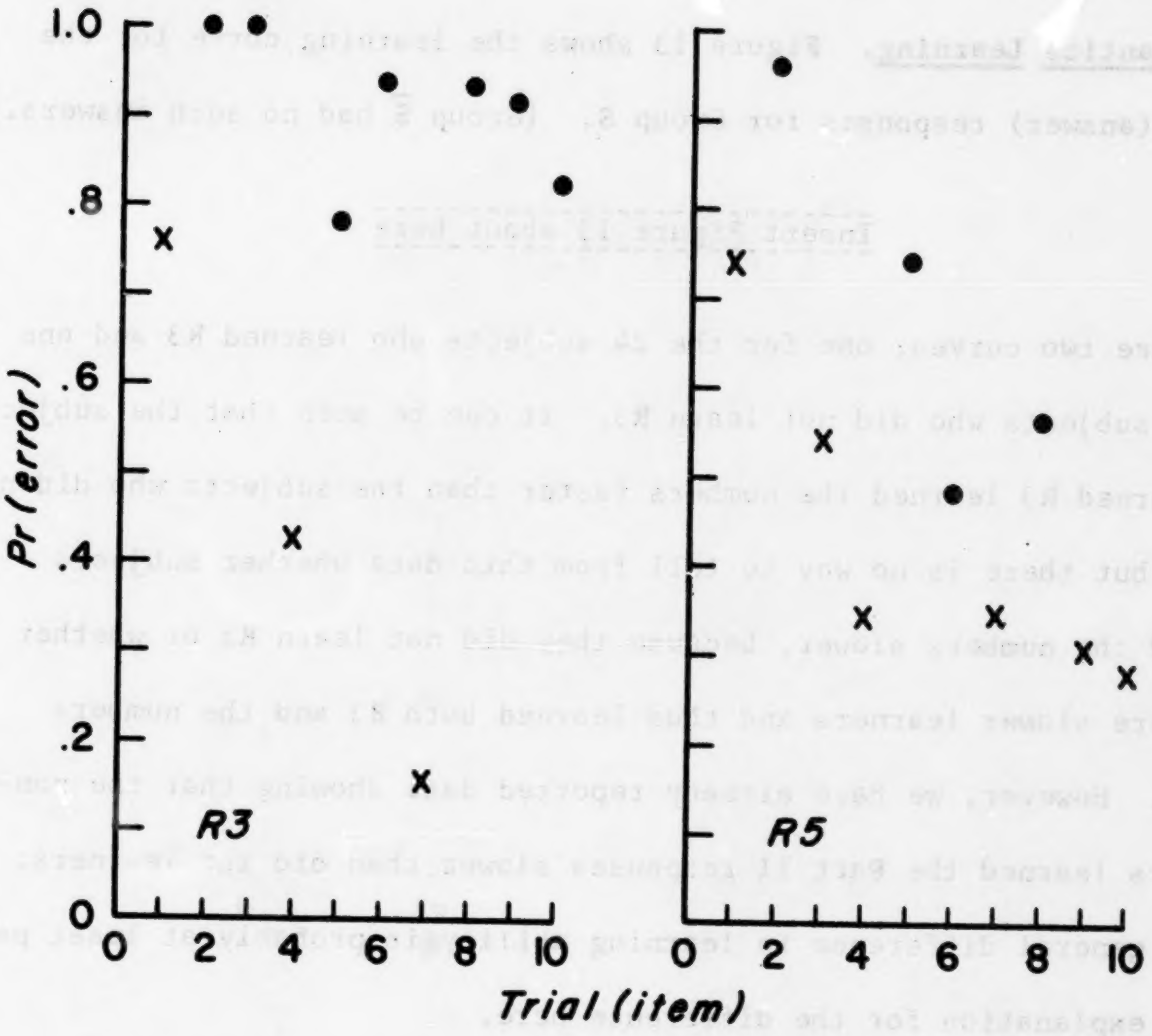

Fig. 12. Learning curves for first 10 trials (items) for R3 and R5. For R3, x's are trials on which S1 = I and dots are trials on which S1 = N. For R5, x's are trials on which S5 = N, and dots are trials on which S5 = 1.

the dots are trials on which S5 = I. Here it is also clear that the curve is not monotonic, the x's representing fewer errors. We may conclude that R2, R4 and R6 were not learned independently on the different kinds of trials.

Semantics Learning. Figure 13 shows the learning curve for the number (answer) responses for Group S. (Group $\overline{S}$ had no such answers.)

Insert Figure 13 about here

There are two curves, one for the 24 subjects who learned R3 and one for the 10 subjects who did not learn R3. It can be seen that the subjects who learned R3 learned the numbers faster than the subjects who did not learn, but there is no way to tell from this data whether subjects learned the numbers slower, because they did not learn R3 or whether they were slower learners and thus learned both R3 and the numbers slower. However, we have already reported data showing that the non-learners learned the Part II responses slower than did the learners. Thus a general difference in learning ability is probably at least part of the explanation for the difference here.

Both groups of subjects approached an asymptote of no errors. So this simple semantic system can be learned quite readily. Since this system is somewhat simpler than the syntactic system discussed earlier, let us look at some of the properties of learning the system. A simple one-element model will not work because inspection of the data reveals that there were more errors on the first few trials, even when trials after the last error were excluded. However, another possibility suggests itself. Many of the responses were wrong because they are sums

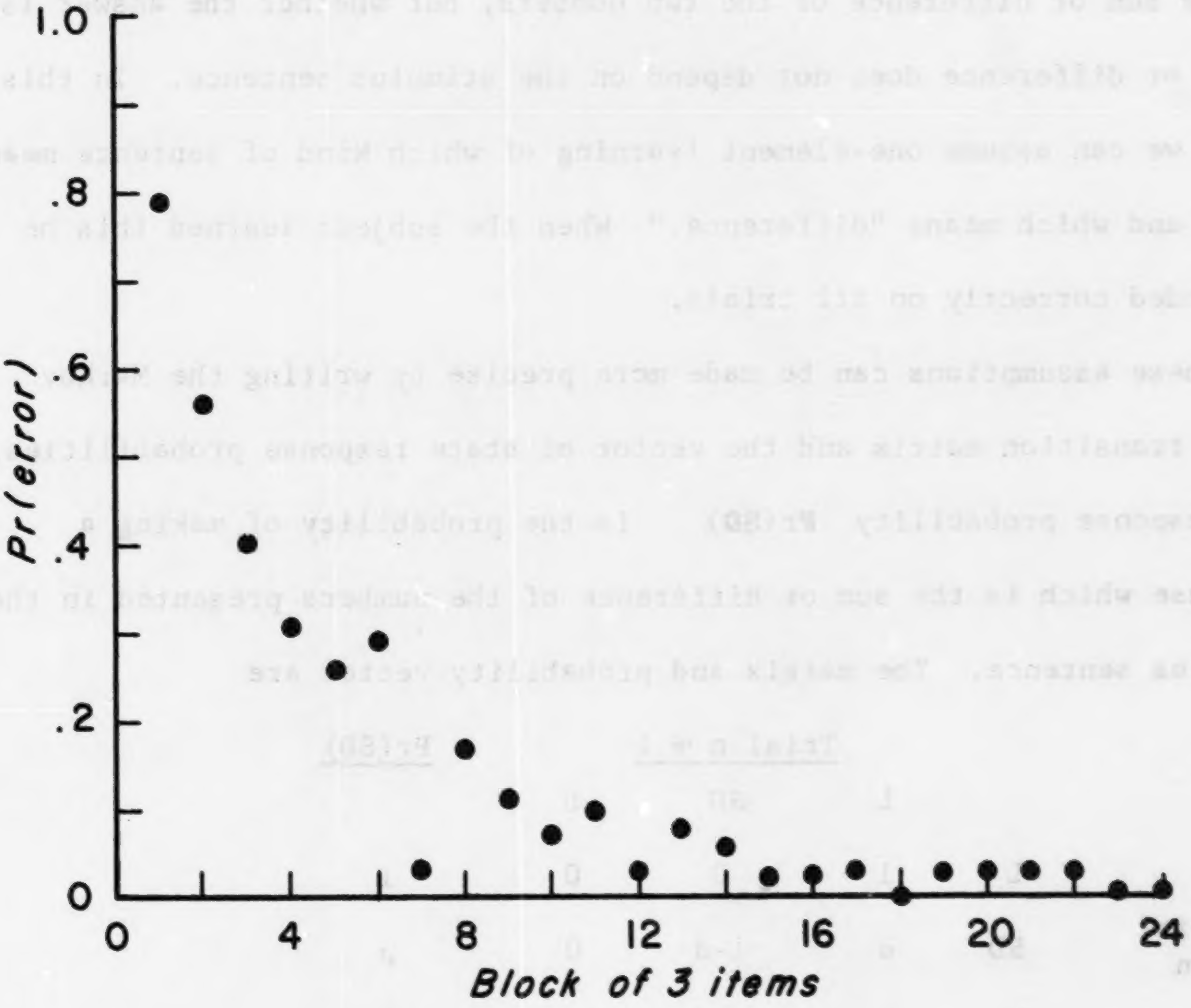

Fig. 13. Learning curve for number (semantics) responses.

of the two numbers in the sentence when they should be differences or vice versa. We assume that at first the subject did not even respond with sums or differences. In this state the subject answered randomly or made no response at all. We can assume one-element learning to take the subject into state SD, where he mostly responed with an answer which is the sum or difference of the two numbers, but whether the answer is a sum or difference does not depend on the stimulus sentence. In this state we can assume one-element learning of which kind of sentence means "sum" and which means "difference." When the subject learned this he responded correctly on all trials.

These assumptions can be made more precise by writing the Markov chain transition matrix and the vector of state response probabilities. The response probability Pr(SD) is the probability of making a response which is the sum or difference of the numbers presented in the stimulus sentence. The matrix and probability vector are

$$\begin{array}{cccccc} & & \multicolumn{3}{c}{\underline{\text{Trial } n + 1}} & \underline{Pr(SD)} \\ & & L & SD & U & \\ & L & 1 & 0 & 0 & 1 \\ \text{Trial } n & SD & d & 1-d & 0 & p \\ & U & 0 & c & 1-c & 0 \end{array}$$

We assumed that in the unlearned state the probability of a subject's making a sum or difference response is 0 even though it might be a little higher than that because when the subject guessed a numeral he might have guessed such a response. However, the probability is quite a bit smaller than 2/10 (there were 10 possible answers, as the subject knew) because many responses in the early part of the response protocols were blank.

It is important to realize that this theory does not distinguish between the two kinds of sentences, i.e., Sum (S) sentences, where the sum of the numbers is correct and Difference (D) sentences, where the difference of the two numbers is correct.

The transition matrix is the same as for some cases of the two-element model (e.g., Bower and Theois, 1964). We attempted to estimate parameters for the above model by applying the methods of Greeno (1968). This analysis was more appropriate than other analyses because it allowed subjects to start in a state other than the unlearned state. Since some subjects were correct on the first trial this was necessary. Greeno's Case 2 analysis was applied, which was the natural one for our data. The theory was applied to the 24 subjects who learned R3, using Greeno's matched-statistics estimates for parameters. However, no matter what identifying restriction was assumed (i.e., learning on correct or error trials out of the intermediate state is equivalent, or there are no transitions to the learned state from the unlearned state), the estimates were not acceptable, some of them either being negative or greater than one. The problem is that we have too little data for making reliable estimates of statistics, there being only 24 learning sequences. For example, an important statistic in the estimation method is the number of errors before the first correct response made by subjects who made no errors after the first correct response. However, there were only four such subjects in our data, and thus, the estimate could not be considered reliable. Since these methods just did not work, there is no reason to analyze them further. If we were interested primarily in this question, an experiment could be arranged which would allow a better test of the model.

One prediction from such a model is the following. If in the unlearned state the subject never makes a sum or difference response, and if in the intermediate state he makes such a response with constant probability, then the plot of proportion of errors versus trials after the first sum or difference response, for responses before the last error, should be horizontal. Figure 14 shows this plot. It looks roughly flat, though

Insert Figure 14 about here

we have left out trials at the end where there were only a few subjects. χ^2 (between theory and data) = 3.04, 4df, $p > .50$. A t test of the difference between the number of errors in the first half and second half of a subject's protocol (responses after first correct and before last error) is significant ($t = 2.14$, 23 df, $p < .05$), more errors occurring in the second half. However, the significance is due to a small variance, the mean numbers of errors for the two halves differing by less than 1.

The model makes another prediction, a prediction which relates specifically to the difference and sum sentences. Let Pr(S/D) be the probability of giving a sum response to a difference sentence, and define the other three probabilities likewise. Then the model predicts that in State SD, Pr(S/D) = Pr(D/D) and Pr(S/S) = Pr(D/S). Once again we look at trials on which we know subjects were in state SD, those after the first sum or difference response and before the last error. Table 9 shows the above probabilities for these trials. We see that the

Insert Table 9 about here

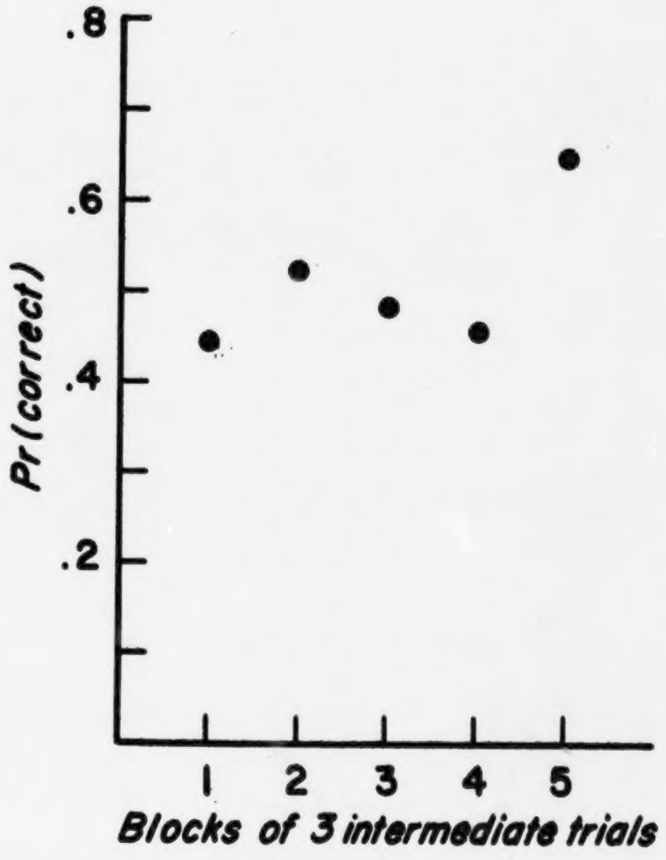

Fig. 14. Stationarity curve for number responses after first sum or difference response and before last error.

TABLE 9

Probability of Giving a Sum or Difference Response to a Sum or Difference Sentence. Only Trials after the First Sum or Difference Response and before the Last Error are Included.

		Response	
		Sum	Difference
Stimulus Sentence	Sum	.53	.06
	Difference	.35	.37

model is wrong in this prediction. The subjects are much more likely to give a difference response to a difference sentence than to a sum sentence. Somehow the subjects have some knowledge about sum sentences and do not give difference responses to them.

Part IV - Grammaticality Learning

There can be two kinds of errors in Part IV, either a 1 response where a 0 was correct (i.e., calling the sentence grammatical when it was ungrammatical) or a 0 where a 1 was correct (calling the sentence ungrammatical when it was grammatical). For now we consider both kinds together and simply call them errors. Figure 15 shows the learning curves for Part IV for the 50 subjects who learned R3 and for the 23 subjects who did not learn R3. Excluded from the curve is trial number 16, because the reading of the sentence was garbled. The number of errors for this response was higher than for the responses adjacent to it, but this was doubtless due to the lack of clarity of the sentence. For each trial, whether the sentence was grammatical (1) or ungrammatical (0) is indicated at the bottom of the figure. Asterisks indicate the four special ungrammatical sentences in which sentence words were interchanged.

Insert Figure 15 about here

First we see that, as a group, subjects who learned R3 also learned Part IV. The mean number of errors per subject per trial over the last 6 trials is .03. If the subjects guessed 0 or 1 with probability ½ each, the mean would be .50. Did the subjects start Part IV always being correct? Since they had learned R3 by definition (that is, by selection of subjects) and since, according to the results discussed for Part III,

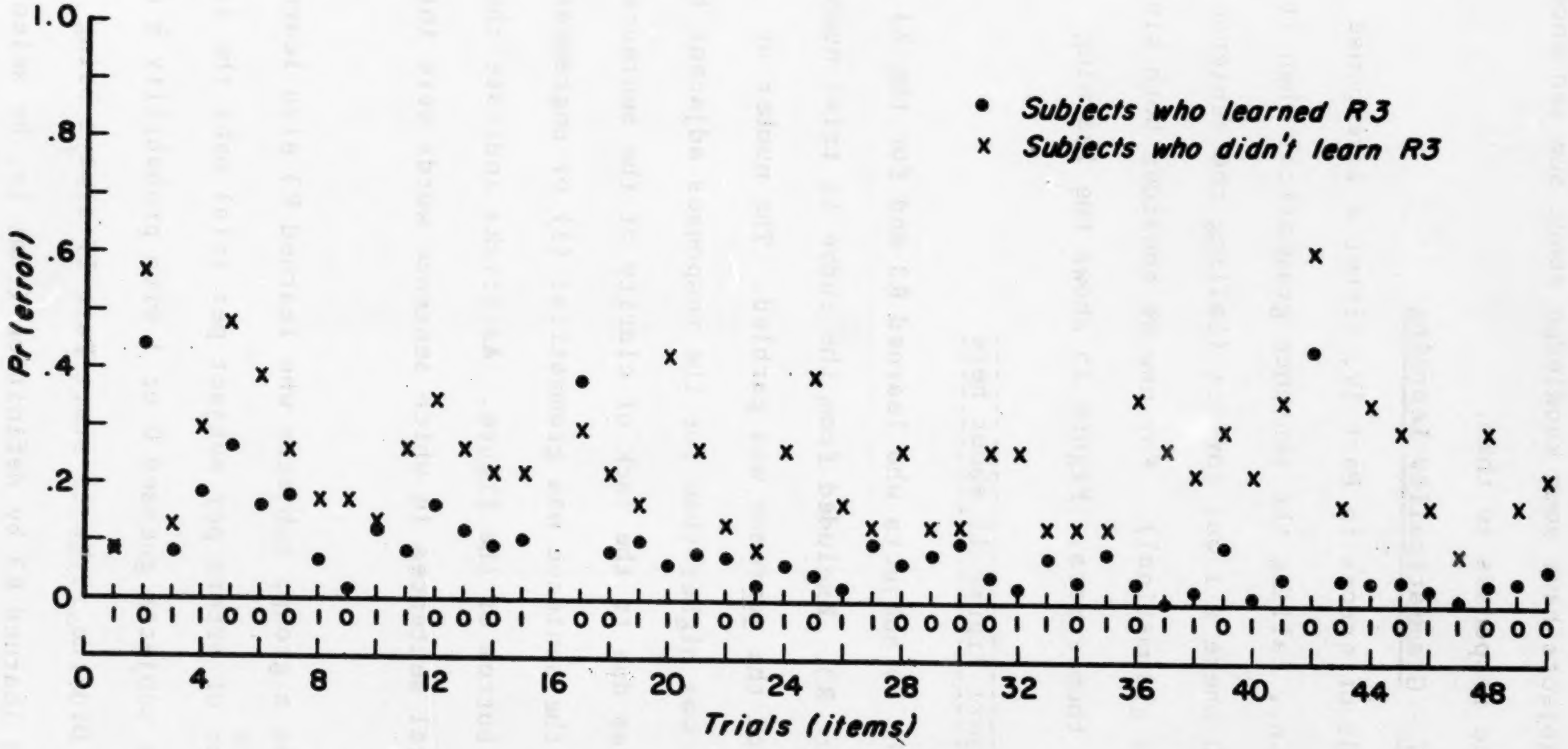

Fig. 15. Learning curve for Part IV, judgments of grammaticality. 0 indicates ungrammatical sentence, 1 grammatical. Asterisk (*) indicates 4 special ungrammatical sentences.

they also learned R1 and R5, it is possible that they could have done perfectly on Part IV from the start. That is, because the response rule for R1, R3, and R5 could have been coded as "the stimulus is always an I or N, and there is exactly one I," the subjects might have used this to respond correctly on Part IV.

But it is clear that the subjects did not start out by almost always being correct. The proportion of errors on trial 1 was only .08, but on trial 2 it shot up to 0.44. Note that on trial 1 a grammatical sentence was presented but on trial 2 an ungrammatical sentence was the stimulus. Since the proportion of errors on trial 1 is only .08, it seems clear that subjects did not guess 0 and 1 each with probability ½. But could they be simply guessing 1 with probability close to 1? No, because then the proportion of errors on trial 2 would be close to 1, instead of .44.

The question is, do subjects recognize at first that a sentence with "ikutsu" appearing twice (i.e., an ungrammatical sentence) is different from one that has only one "ikutsu?" If they did not distinguish between them, the proportion of errors on trials 1 and 2 would not be different (i.e., if the subjects were guessing independently of the stimulus sentence, no matter what the guessing probability, the expected proportions of errors on the two trials would be the same. This assumes, of course, that no learning occurs between the first and second trials. But there seems no reason to suppose that learning to distinguish between a G sentence and a U sentence would occur as the result of one exposure to a G sentence. And if learning did occur, the proportion of errors for trial 2 would be lower than for trial 1, not higher, which was the actual result). Therefore, it seems likely that from the start subjects discriminated the ungrammatical from the grammatical sentences, but had to learn how to respond to them.

Bearing these results in mind, let us look at the results for the 23 subjects who did not learn R3. Table 10 shows the mean number of errors per subject for both groups (i.e., those who learned or did not learn R3). The number of errors is greater for the group that did not learn R3 than for the group that did learn. This is a result we would expect, since if a subject did not learn R3 we might assume he had not learned that an I could not appear twice. But suppose we assume that the subject had learned nothing about this. Once again this would lead us to predict that the proportions of errors for trials 1 and 2 would be the same. Figure 15, however, shows this is not the case; the proportions are .09 and .57, respectively. These proportions are not way out of line with the proportions for subjects who learned R3. The best explanation for this result seems to be that even subjects who did not learn R3 by our definition learned the structure of the syntax, i.e., that I appeared exactly once. Remember that many of the subjects in this group had never used two responses in a box, i.e., they had not followed directions. Also, only three subjects had locked into an R3 response of N,I. What seems to have happened then is that even most of the 23 subjects in this group learned the structure or something about the structure, which leads to the different proportions between trials 1 and 2.

Insert Table 10 about here

Table 11 shows the number of subjects in each group who made at least one error on the last 6 trials. Consistent with the results we

Insert Table 11 about here

TABLE 10

Mean Number of Errors in Part IV, Grammaticality Judgments.

				Total	
	$\overline{SW}$	$\overline{S}W$	$\overline{SA}$	$\overline{S}$	S
Learned R3	4.8	2.8	2.9	3.4	6.5
Did Not Learn R3	13.8	4.0	18.5	14.5	10.3

TABLE 11

Proportion of Subjects who Made at Least One Error on Last 6 Items of Part IV, Grammatical Judgments.

	S	$\overline{S}$	Total
Learned R3	.17	.04	.10
Did Not Learn R3	.30	.62	.48

have already discussed, the subjects who did not learn R3 had proportionately higher scores on this statistic than subjects who did learn. In fact, 11 of the 23 non-learning subjects made at least one error on the last 6 trials. It is possible that some subjects who did not learn R3 because of lack of ability or motivation had the same effect on Part IV. This is substantiated by the fact that these subjects also did less well on Part II.

Now let us turn to the four special ungrammatical sentences. We can read the proportion of errors for each from Figure 15. The first of these sentences was presented on trial 9 and read "ikutsu desuka 1 wa 3 tasu." In other words, "tasu" and "desuka" were interchanged. Considering the 50 subjects who learned R3, only a proportion of .02 of them called this sentence grammatical. The second of these sentences appeared on trial 17 and read "0 5 tasu wa ikutsu desuka." In other words, "tasu" and "5" were interchanged. The proportion of errors was .38. This proportion was much higher than the proportion for the trial immediately preceding and following it. The third sentence was number 30 and was supposed to have read "2 ikutsu tasu wa 3 desuka." In other words, "ikutsu" and "tasu" were interchanged. However, the speaker made an error and instead of saying "tasu" he said something that sounded like "des." In other words, a new word was introduced to the subjects. The proportion calling this sentence grammatical was only .10. However, this proportion was doubtless low because of the introduction of the new word, so we will not consider this sentence. The fourth sentence was number 42 and read "ikutsu wa 2 tasu 4 desuka." Here "wa" and "tasu" were interchanged. The proportion of subjects responding 1 (grammatical) was .44. Once again, this proportion was much larger than for the sentences immediately preceding and following it.

The question that strikes us is, why is the proportion of errors so much higher for sentences 17 and 42 (.38 and .44) than for sentence 9 (.02)? Two explanations suggest themselves. First, consider the "word distance" between the two words interchanged to make the ungrammatical sentence from a grammatical sentence. This is 1 plus the number of words between the two words in the grammatical sentence. This measure for the three sentences is: for sentence 9 the distance is 4, for sentence 7 the distance is 1, and for sentence 42 the distance is 2. So it is a question of distance 4 on the one hand versus distances 1 and 2 on the other. It might be that this distance is a good measure of sentence grammaticality. The greater the distance the more chance the sentence will be called ungrammatical.

However, another possibility is that sentence 9 was heard as ungrammatical because it put "desuka" out of place. "Desuka" is the last word of every sentence and signals the subject that the sentence is over. When it did not appear there, but tasu appeared in its place, this was probably very salient to the subject. As we saw previously, "desuka" (R6) was the response learned quickest in Part III. This was doubtless not because of the properties of the word, but because it appeared last.

There is no way to distinguish in this experiment between these two possibilities. An experiment could be done varying this "word distance" and having subjects judge grammaticality. However, there does seem to be one solid conclusion from the results. That is that subjects make many more errors in this part on the few sentences which interchange function words than on the sentences which include "ikutsu" twice. Whether this is due to more practice on the latter or to some other reason is not clear.

In summary, Part IV mainly confirmed our belief that on Part III, subjects learned the language J. It has also provided evidence that some subjects who did not learn R3 by our definition did indeed learn the language J.

VI. Discussion and Summary

The major point of our study was to try to decide what kind of automaton best represents a subject's behavior in the experiment. First, we noted that if the subject became an ordered-state finite automaton, he would not learn the syntax of J. The results presented in the last section show that most of the subjects who followed the instructions learned, and that of those who did not, only three behaved at asymptote in the way a sequential finite automaton such as $\mathcal{A}$ might predict. Also, the results of Part IV of the experiment suggested that even the nine subjects who did not learn R3 by our criteria learned much of the structure of J. We may safely conclude that, in general, subjects did not behave as if they became ordered-state finite automata.

We predicted that if subjects became either general finite automata or ordered-state 1-memory store automata, then they would learn, as they could become either $\mathcal{J}$ or $\mathcal{P}$. However, we noted a way to distinguish between these two automata. By making a general assumption about the course of learning on finite automata and 1-MS automata, we could write equations (1) in Section II for $\mathcal{J}$ and equations (2) for $\mathcal{P}$. The equations for the finite automaton $\mathcal{J}$ predict that R2 is learned at the same rate as R1, while the equations for $\mathcal{P}$ predict that R2 is learned faster than R1. By the same reasoning that produced these equations, we can derive similar equations which predict for $\mathcal{J}$ that R4 is learned at the same rate as R1 and for $\mathcal{J}$ that R4 is learned faster. From both $\mathcal{J}$ and $\mathcal{P}$ we predict that R6 is learned faster than R1. The difference between R2 and R6 here is that in $\mathcal{J}$ the pair (s_9, D) appears on every trial. It is clear that we cannot write a finite automaton that will behave like R6 for all responses, including R2 and R4, since then we would have an ordered-state finite automaton, and we saw in Section II that no ordered-state finite automaton can respond correctly to J.

Now, the above predictions are made with respect to R1. But by exactly the same reasoning, we see that $\mathcal{J}$ predicts that R3 and R5 are learned at the same rate as R2 and R4, while $\mathcal{P}$ predicts that R2 and R4 are learned faster. Both automata predict that R6 is learned faster than R3 or R5. In short, $\mathcal{J}$ predicts that R1 through R5 are learned at the same rate, while $\mathcal{P}$ predicts that the even responses (R2,R4) are learned faster than the odd responses.

We saw in the last section that, in fact, no matter what statistic we looked at, all the even responses were learned faster than the odd responses, and this result even held across the four sub-groups. These results make it clear that the predictions from $\mathcal{P}$ are much closer to the experimental data than are the predictions from $\mathcal{J}$. In this experiment, at least, subjects behaved more like a 1-MS than like a finite automaton.

An alternative explanation of our results might be proposed. This is that, for some reason, it is difficult for the subject to learn those responses where a two-letter response is correct. This would explain why R1 and R3 were learned slowly compared with the even responses, but it would not explain why R5 was learned more slowly than the even responses, because the correct responses for R5 contained one letter (N or I depending on the history). This built-in control rules out the two-letter explanation.

Also, note that the usual serial position effect could not explain our results. The results do not at all fit a bowed serial position curve (where the serial position is R1 through R6). In fact, an error curve through the results (as well as predictions) changes its direction (i.e., the sign of the first derivative) at every point. For example, there are more errors for R3 than for R2 or R4, and this could not occur in a bowed serial position curve.

In addition to the above predictions, as we saw in Section II, our learning assumption together with $\mathcal{J}$ predicts that each kind of trial is the same. Specifically, $\mathcal{J}$ predicts that the learning curve over all the trials should be monotonically decreasing, but that the points for one of the kinds of trials should come up. We saw in the last section that the curves were monotonically decreasing for both even responses. Once again, the 1-MS $\mathcal{P}$ is more appropriate for the data.

We also wanted to look at the effects of semantics practice on the learning of syntax. The hypothesis that semantics acts as a motivator only predicts that the semantics group would do better than the non-semantics group on all the responses. The hypothesis that semantic structure restricts the range of possible syntactic structures predicts that, since this restriction only affects R3 and R5 (since these responses are the only ones affected by the history of the sequence), the semantics group would do better on these responses, but there would be no difference on the other responses between the two groups.

The results show that indeed there was no difference on mean trial of last error between the two groups on R1, R2, R4 and R6, as the restriction hypothesis suggests, and that the semantics group did better on R5, again as the restriction hypothesis suggests. On the other hand, R3 was not significantly better for the semantics group. However, the mean was smaller for the semantics group on R3, and this was the only response besides R5 for which this was true. At any rate, since R5 was the one response for which the semantics group did significantly better, these results, though less conclusive than our results on the syntax learning, suggest that the restriction hypothesis predicts the data better than does the motivation hypothesis.

We also saw that the semantics system (correct number responses) was learned by the subjects. There was some evidence that before the subjects were in a state in which they always answered correctly, they were guessing numbers which were sums or differences of the two numbers presented in the sentence.

Do our results suggest anything about language learning in general? It is of some interest that a finite automaton did not turn out to be an appropriate representation for the subject in our experiment. Of course, the language we dealt with was a finite language so that it is not a question of generative capacity. Our 1-MS is much weaker than the general PDS automata. On the other hand, a crucial part of the PDS structure remains in our version and distinguishes it from finite automata. This structure is that there is memory besides the state of the automaton. Perhaps our experimental results are generalizable to more complex languages, including languages with loops, which we have not considered at all in this study.

Our results on semantics suggest that studies of syntax learning that do not include a semantic model may be losing an important component of syntax learning. The results seem to suggest that semantics acts as more than a motivator.

In general, we feel that the value of our study lies in the fact that it provided experimental evidence for the kind of automaton a person could become. The predictions from the automata included both predictions about whether a person who became a given kind of automaton could learn a given language, and also predictions about how a language would be learned. These predictions allowed us to distinguish between various kinds of automata. Perhaps future work on more complex languages will confirm our results.

Appendix I. Transformational Rules for Arithmetic

Our purpose is to list the transformational rules for a subset of spoken arithmetic in English and Japanese. We do not give any discussion of the rules. Our goal is mainly to show that spoken arithmetic can be generated by a miniature linguistic model having the properties of the model discussed in Section III.

Notation is the standard linguistic one. All transformations (except the lexical ones) are described by an analysis, which is a cut of the phrase-marker of a sentence, and a permutation of that analysis. For each transformation, we call the analysis A and the permutation P. When we write BLOCK, it is the same as writing the empty string, but we do it this way for graphic purposes. The transformations are ordered and, except for those labelled otherwise, are obligatory. The transformations apply to the base in Table 1.

The BLOCK transformations are used to delete strings that do not have the proper number of x's (variables) for the given sentence. This is related to the discussion of base strings whose meaning is empty in Section III. However, some strings are deleted whose meaning is not empty, namely, strings with more than one variable, since there is no natural way of asking such questions in the spoken language, especially when the two variables are not adjacent.

In the rules, capital letters X,Y,Z are variables taking strings as arguments. When such a letter appears, any string can be inserted. Small x is the variable in arithmetic.

English Transformations

A. Lexicon

0 → zero
1 → one
2 → two
3 → three
4 → four
5 → five
= → equal
+ → plus
- → minus
· → times
/ → divided by
(→ ∅
) → ∅

B. Sentence Transformations

1. T_{BL1} A = X,x,Y,x,Z
   P = 1 2 3 4 5 → BLOCK
2. $T_{Q_{wh1}}$ A = Q_{wh},X,x,Y,=,Z
   P = 1 2 3 4 5 6 → 2 what 4 is 6
3. $T_{Q_{wh2}}$ A = Q_{wh},X,=,Y,x,Z
   P = 1 2 3 4 5 6 → 2 is 4 what 6
4. T_{BL2} A = Q_{wh},X
   P = 1 2 → BLOCK
5. T_{BL3} A = X,x,Y
   P = 1 2 3 → BLOCK
6. $T_{Q_{YN}}$ A = Q_{YN},X,=,Y
   P = 1 2 3 4 → Does 2 3 4
7. T_{CA} A = C,(,N,+,N,)
   P = 1 2 3 4 5 6 → Add 3 and 5
8. T_{CS} A = C,(,N,-,N,)
   P = 1 2 3 4 5 6 → Subtract 5 from 3
9. T_{CM} A = C,(,N,.,N,)
   P = 1 2 3 4 5 6 → Multiply 3 by 5
10. T_{CD} A = C,(,N,/,N,)
    P = 1 2 3 4 5 6 → Divide 3 by 5

Japanese Transformations

A. Lexicon

0 → zero
1 → ichi
2 → ni
3 → san
4 → shi
5 → go
= → wa
+ → tasu
- → hiku
. → karu
/ → waru
(→ ∅
) → ∅

B. Sentence Transformations

1. T_{BL1} $\quad A = X,x,Y,x,Z$

$$P = 1\ 2\ 3\ 4\ 5 \rightarrow \text{BLOCK}$$

2. $T_{Q_{wh1}}$ $\quad A = Q_{wh},N \left\langle\begin{matrix}+\\-\\.\\/\end{matrix}\right\rangle ,N,=,N$

$$P = 1\ 2\ \ 3\ \ 4\ 5\ 6 \rightarrow 1\ 2 \left\langle\begin{matrix}\text{ni}\\\text{kara}\\\text{ni}\\0\end{matrix}\right\rangle 4 \left\langle\begin{matrix}0\\0\\0\\\text{de}\end{matrix}\right\rangle 3 \text{ to } 6$$

3. $T_{Q_{wh2}}$ $\quad A = Q_{wh},X,x,Y$

$$P = 1\ 2\ 3\ 4 \rightarrow 2 \text{ ikutsu } 4 \text{ desuka}$$

4. T_{BL2} $\quad A = Q_{wh},X$

$$P = 1\ 2 \rightarrow \text{BLOCK}$$

5. T_{BL3} $\quad A = X,x,Y$

$$P = 1\ 2\ 3 \rightarrow \text{BLOCK}$$

6. T_{YN} $\quad A = Q_{YN},X$

$$P = 1\ 2 \rightarrow 2 \text{ desuka}$$

7. T_C $\quad A = C,(,N, \left\langle\begin{matrix}+\\-\\.\\/\end{matrix}\right\rangle ,N,)$

$$P = 1\ 2\ 3\ \ 4\ \ 5\ 6 \rightarrow 3 \left\langle\begin{matrix}\text{ni}\\\text{kara}\\\text{ni}\\0\end{matrix}\right\rangle 5 \left\langle\begin{matrix}0\\0\\0\\\text{de}\end{matrix}\right\rangle 4 \text{ te kudasai}$$

In this last transformation we have ignored a morphophonemic rule that takes, for example, tasu + te → tashite.

Appendix II. Experimental Instructions for Part III, Group S

Part III will probably be more difficult than the other parts. The instructions are somewhat complex, so listen carefully. You are going to learn some simple Japanese sentences. Each sentence contains six words. You are already familiar with all the words. They are all either the four words you became familiar with in Part I or they are the numbers you learned in Part II. Your first job is to learn to predict what the order of words is in each sentence. You will hear a tone (or a bleep) on the television. Then you will write the letters for what you think the first words can be in the first box. If you think the word will be one of the four words you learned in Part I, write the first letter of that word, for example, T for tasu. However, if you think the word will be one of the numbers you learned, write N for number. Remember, do not write the first letter of a particular number, rather write N for number. In some sentences, in some positions, it is possible that more than one word could occupy that position. In fact, sometimes two words could possibly occupy a position. If you think only one word can occupy a position, write the letter for that word before the comma in the box. If you think two words could occupy the position, write both words, one before and one after the comma. Remember, in some sentences, in some positions only one word would be correct, and in some positions two words would be correct. So do not always fill the space after the comma because sometimes only one would be correct. The patterns are such that sometimes a preceding word can influence what words can later follow. So do not write all six answers at one time. Always fill in just one box, then wait for the next word to be spoken. You have a few seconds to make your prediction. Then the actual word of the sentence will be said by the speaker. This may be only one of the possible words that might appear at that position. If you predicted this word you were correct. If you predicted another word you might have been correct. Since at most two words could have come in that position, if you predicted two words and neither was said by the speaker, at least one of them was wrong.

After you hear the first word of the sentence there will be a few seconds' pause and you will then predict the second word of the sentence. Then the third word will come, and so on, for the six words. Please do not write any answers after you have heard the correct word. We have to trust you, and it is very important to us to get your answers before you have actually heard the correct answer. Look at your answer sheets. Each row is for one sentence. The row of six boxes is for the six predictions of the words in the sentences. The comma is there so that you may predict two words if you wish. Please predict only words that you feel might be correct. If you have some feeling that they are correct, write them. But do not make completely wild guesses. If you do not know any word you want to predict, put a dash in the box and write the next answer in the next box. Are there any questions about this part of the procedure?

There is one thing more to this part. Please listen carefully. Each of these sentences is an actual Japanese sentence. And each one is a sentence asking a question in arithmetic. The questions are about addition. In algebra the questions they ask would be expressed by the equations, for example, "1 plus 3 equals x," "1 plus x equals 3," and "x plus 1 equals 3." That is, the required answer is the value of x.

These are the only sentences you will be hearing. In English, the questions would be, perhaps, "1 plus 3 equals what," "1 plus what equals 3," and "what plus 1 equals 3?" Note that the answer to, say, "1 plus what equals 3" is "2" whereas the answer to "1 plus 3 equals what" is "4." That is, the answers are different. It is also your job in this part to learn the meaning of these Japanese sentences, that is, to learn what questions the sentences are asking. Remember, the sentences all have the meaning of one of the 3 algebraic equations I mentioned before. After you have heard the six words of each sentence repeated slowly, and you have made your predictions, you will hear the same sentence, repeated at a more natural speed. Then you have a few seconds to write the answer to that sentence in the box to the right of the six boxes and separated from it. Then the numerical answer will appear on the screen. For example, if you think the sentence asked the question (in Japanese), "x plus 1 equals 3," the number 2 will appear. If the question is "1 plus 3 equals x," the number 4 will appear. Once again, please do not write any answers after you have seen the correct answer. If you do not know an answer put a dash in the box. Do not try to write the Japanese number for these answers. Simply write the digit. The answers are any number from 0 to 9. After the numerical answer appears on the screen, a tone will once again be heard. This is your signal to predict the first word of the next sentence. Are there any questions?

References

Bower, G.H., & Theios, J. A learning model for discrete performance levels. In R.C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford Univ. Press, 1964, pp. 1-31.

Braine, M.D.S. On learning the grammatical order of words. *Psychol. Rev.*, 1963, *70*, 323-348.

Chomsky, N. *Aspects of the theory of syntax*. Cambridge, Mass.: M.I.T. Press, 1965.

Chomsky, N. Formal properties of grammars. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. II. New York: Wiley, 1963, pp. 323-418.

Chomsky, N. & Miller, G.A. Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. II. New York: Wiley, 1963, pp. 269-322.

Crothers, E.J. & Suppes, P. *Experiments in second-language learning*. New York: Academic Press, 1967.

Epstein, W. The influence of syntactic structure on learning. *Amer. J. Psychol.*, 1962, *74*, pp. 80-85.

Ervin-Tripp, S. Changes with age in the verbal determinant of word-association. *Amer. J. Psychol.*, 1961, *74*, pp. 361-372.

Ginsburg, S. *An introduction to mathematical machine theory*. Reading, Mass.: Addison-Wesley, 1962.

Ginsburg, S. *The mathematical theory of context-free languages*. New York: McGraw-Hill, 1966.

Greeno, J.G. Identifiability and statistical properties of two-stage learning with no successes in the initial stage. *Psychometrika*, 1968, *43*, pp. 173-215.

Katz, J.J. & Postal, P. *An integrated theory of linguistic descriptions*. Cambridge, Mass: M.I.T. Press, 1964.

Lakoff, G. & Ross, J.R. Is deep structure necessary? Duplicated. Cambridge, Mass.: M.I.T. Dept. of Linguistics, 1967.

Miller, G.A. & Norman, D. Research on the use of formal languages in the behavioral sciences. Technical Report, January-June 1964, Department of Defense, Advanced Research Projects Agency, Harvard University, 1964.

Minsky, M. Introduction. In M. Minsky (Ed.), Semantic information processing. Cambridge, Mass.: M.I.T. Press, 1968, pp. 1-32.

Minsky, M. (Ed.). Semantic information processing. Cambridge, Mass.: M.I.T. Press, 1968.

Newell, A., Shaw, J.C. & Simon, H.A. Report on a general problem-solving program. Information Processing. Proc. International Conference on Information Processing. Paris: UNESCO, 1959, pp. 256-264.

Rabin, M.O. & Scott, D. Finite automata and their decision problems. IBM Journal of Research and Development, 1959, 3, pp. 114-125.

Suppes, P. Stimulus-response theory of finite automata. Tech. Rep. No. 133, Institute for Mathematical Studies in the Social Sciences, Stanford University. Published in Journal of Mathematical Psychology, 1969, 6, 327-355.

(Continued from inside front cover)

96 R. C. Atkinson, J. W. Breisford, and R. M. Shiffrin. Multi-process models for memory with applications to a continuous presentation task. April 13, 1966. (*J. math. Psychol.*, 1967, *4*, 277-300).

97 P. Suppes and E. Crothers. Some remarks on stimulus-response theories of language learning. June 12, 1966.

98 R. Bjork. All-or-none subprocesses in the learning of complex sequences. (*J. math. Psychol.*, 1968, *1*, 182-195).

99 E. Gammon. The statistical determination of linguistic units. July 1, 1966.

100 P. Suppes, L. Hyman, and M. Jerman. Linear structural models for response and latency performance in arithmetic. In J. P. Hill (ed.), *Minnesota Symposia on Child Psychology*. Minneapolis, Minn.: 1967. Pp. 160-200).

101 J. L. Young. Effects of intervals between reinforcements and test trials in paired-associate learning. August 1, 1966.

102 H. A. Wilson. An investigation of linguistic unit size in memory processes. August 3, 1966.

103 J. T. Townsend. Choice behavior in a cued-recognition task. August 8, 1966.

104 W. H. Batchelder. A mathematical analysis of multi-level verbal learning. August 9, 1966.

105 H. A. Taylor. The observing response in a cued psychophysical task. August 10, 1966.

106 R. A. Bjork. Learning and short-term retention of paired associates in relation to specific sequences of interpresentation intervals. August 11, 1966.

107 R. C. Atkinson and R. M. Shiffrin. Some Two-process models for memory. September 30, 1966.

108 P. Suppes and C. Ihrke. Accelerated program in elementary-school mathematics--the third year. January 30, 1967.

109 P. Suppes and I. Rosenthal-Hill. Concept formation by kindergarten children in a card-sorting task. February 27, 1967.

110 R. C. Atkinson and R. M. Shiffrin. Human memory: a proposed system and its control processes. March 21, 1967.

111 Theodore S. Rodgers. Linguistic considerations in the design of the Stanford computer-based curriculum in initial reading. June 1, 1967.

112 Jack M. Knutson. Spelling drills using a computer-assisted instructional system. June 30, 1967.

113 R. C. Atkinson. Instruction in initial reading under computer control: the Stanford Project. July 14, 1967.

114 J. W. Breisford, Jr. and R. C. Atkinson. Recall of paired-associates as a function of overt and covert rehearsal procedures. July 21, 1967.

115 J. H. Stelzer. Some results concerning subjective probability structures with semiorders. August 1, 1967

116 D. E. Rumelhart. The effects of interpresentation intervals on performance in a continuous paired-associate task. August 11, 1967.

117 E. J. Fishman, L. Keller, and R. E. Atkinson. Massed vs. distributed practice in computerized spelling drills. August 18, 1967.

118 G. J. Groen. An investigation of some counting algorithms for simple addition problems. August 21, 1967.

119 H. A. Wilson and R. C. Atkinson. Computer-based instruction in initial reading: a progress report on the Stanford Project. August 25, 1967.

120 F. S. Roberts and P. Suppes. Some problems in the geometry of visual perception. August 31, 1967. (*Synthese*, 1967, *17*, 173-201)

121 D. Jamison. Bayesian decisions under total and partial ignorance. D. Jamison and J. Kozielecki. Subjective probabilities under total uncertainty. September 4, 1967.

122 R. C. Atkinson. Computerized instruction and the learning process. September 15, 1967.

123 W. K. Estes. Outline of a theory of punishment. October 1, 1967.

124 T. S. Rodgers. Measuring vocabulary difficulty: An analysis of item variables in learning Russian-English and Japanese-English vocabulary parts. December 18, 1967.

125 W. K. Estes. Reinforcement in human learning. December 20, 1967.

126 G. L. Wolford, D. L. Wessel, W. K. Estes. Further evidence concerning scanning and sampling assumptions of visual detection models. January 31, 1968.

127 R. C. Atkinson and R. M. Shiffrin. Some speculations on storage and retrieval processes in long-term memory. February 2, 1968.

128 John Holmgren. Visual detection with imperfect recognition. March 29, 1968.

129 Lucille B. Mlodnosky. The Frostig and the Bender Gestalt as predictors of reading achievement. April 12, 1968.

130 P. Suppes. Some theoretical models for mathematics learning. April 15, 1968. (*Journal of Research and Development in Education*, 1967, *1*, 5-22)

131 G. M. Olson. Learning and retention in a continuous recognition task. May 15, 1968.

132 Ruth Norene Hartley. An investigation of list types and cues to facilitate initial reading vocabulary acquisition. May 29, 1968.

133 P. Suppes. Stimulus-response theory of finite automata. June 19, 1968.

134 N. Moler and P. Suppes. Quantifier-free axioms for constructive plane geometry. June 20, 1968. (In J. C. H. Gerretsen and F. Oort (Eds.), *Compositio Mathematica*. Vol. 20. Groningen, The Netherlands: Wolters-Noordhoff, 1968. Pp. 143-152.)

135 W. K. Estes and D. P. Horst. Latency as a function of number or response alternatives in paired-associate learning. July 1, 1968.

136 M. Schlag-Rey and P. Suppes. High-order dimensions in concept identification. July 2, 1968. (*Psychom. Sci.*, 1968, *11*, 141-142)

137 R. M. Shiffrin. Search and retrieval processes in long-term memory. August 15, 1968.

138 R. D. Freund, G. R. Loftus, and R.C. Atkinson. Applications of multiprocess models for memory to continuous recognition tasks. December 18, 1968.

139 R. C. Atkinson. Information delay in human learning. December 18, 1968.

140 R. C. Atkinson, J. E. Holmgren, and J. F. Juola. Processing time as influenced by the number of elements in the visual display. March 14, 1969.

141 P. Suppes, E. F. Loftus, and M. Jerman. Problem-solving on a computer-based teletype. March 25, 1969.

142 P. Suppes and Mona Morningstar. Evaluation of three computer-assisted instruction programs. May 2, 1969.

143 P. Suppes. On the problems of using mathematics in the development of the social sciences. May 12, 1969.

144 Z. Domotor. Probabilistic relational structures and their applications. May 14, 1969.

145 R. C. Atkinson and T. D. Wickens. Human memory and the concept of reinforcement. May 20, 1969.

146 R. J. Titiev. Some model-theoretic results in measurement theory. May 22, 1969.

147. P. Suppes. Measurement: Problems of theory and application. June 12, 1969.

148 P. Suppes and C. Ihrke. Accelerated program in elementary-school mathematics--the fourth year. August 7, 1969.

149 D. Rundus and R.C. Atkinson. Rehearsal in free recall: A procedure for direct observation. August 12, 1969.

150 P. Suppes and S. Feldman. Young children's comprehension of logical connectives. October 15, 1969.

(Continued on back cover)

(Continued from inside back cover)

151 Joaquim H. Laubsch. An adaptive teaching system for optimal item allocation. November 14, 1969.
152 Roberta L. Klatzky and Richard C. Atkinson. Memory scans based on alternative test stimulus representations. November 25, 1969.
153 John E. Holmgren. Response latency as an indicant of information processing in visual search tasks. March 16, 1970.
154 Patrick Suppes. Probabilistic grammars for natural languages. May 15, 1970.
155 E. Gammon. A syntactical analysis of some first-grade readers. June 22, 1970.
156 Kenneth N. Wexler. An automaton analysis of the learning of a miniature system of Japanese. July 24, 1970.
157 R. C. Atkinson and J. A. Paulson. An approach to the psychology of instruction. August 14, 1970.